UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MATEUS QUEIROZ CUNHA

UTILIZANDO MODELOS DE LINGUAGEM GRANDES PARA CLASSIFICAÇÃO DE ATOS DO DIÁRIO OFICIAL DA UNIÃO NO DOMÍNIO TRIBUTÁRIO

CAMPINA GRANDE - PB
2024

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Utilizando Modelos de Linguagem Grandes para Classificação de Atos do Diário Oficial da União no Domínio Tributário

Mateus Queiroz Cunha

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Processamento de Linguagem Natural

Prof. Cláudio de Souza Baptista, Ph.D.
(Orientador)

Campina Grande, Paraíba, Brasil

C972u Cunha, Mateus Queiroz.
Utilizando modelos de linguagem grandes para classificação de atos do Diário Oficial da União no domínio tributário / Mateus Queiroz Cunha – Campina Grande, 2024.
146 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.
"Orientação: Prof. Dr. Cláudio de Souza Baptista."
Referências.

1. Processamento de Linguagem Natural. 2. Classificação de Textos. 3. Modelos de Linguagem Grandes. 4. Dados Desbalanceados. 5. Domínio Jurídico. 6. Diários Oficiais. 7. Domínio Tributário. I. Baptista, Cláudio de Souza. II. Título.

CDU 004.43(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA ITAPUANA SOARES DIAS GONÇALVES CRB-15/093

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secretaria-copin@computacao.ufcg.edu.br / copin@copin.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

MATEUS QUEIROZ CUNHA

UTILIZANDO MODELOS DE LINGUAGEM GRANDES PARA CLASSIFICAÇÃO DE ATOS DO DIÁRIO OFICIAL DA UNIÃO NO DOMÍNIO TRIBUTÁRIO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 19/02/2024

Prof. Dr. CLÁUDIO DE SOUZA BAPTISTA, UFCG, Orientador

Profa. Dra. JOSEANA MACÊDO FECHINE RÉGIS DE ARAÚJO, UFCG, Examinadora Interna

Prof. Dr. JOÃO DALLYSON SOUSA DE ALMEIDA, UFMA, Examinador Externo

Prof. Dr. LUCIANO DE ANDRADE BARBOSA, UFPE, Examinador Externo

Documento assinado eletronicamente por **CLAUDIO DE SOUZA BAPTISTA**, **PROFESSOR 3 GRAU**, em 20/02/2024, às 15:43, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **JOSEANA MACEDO FECHINE**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 20/02/2024, às 16:45, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Luciano de Andrade Barbosa**, **Usuário Externo**, em 23/02/2024, às 09:58, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **4205046** e o código CRC **1A001180**.

Referência: Processo nº 23096.008718/2024-11 SEI nº 4205046

Resumo

A Ciência Jurídica destaca-se como um campo promissor para o Processamento de Linguagem Natural, contendo informações relevantes em diversos domínios que impactam a sociedade. O presente estudo concentra-se na identificação de publicações tributárias no Diário Oficial da União (DOU) por meio de uma abordagem de classificação de texto. Durante a análise do contexto tributário no DOU, evidenciou-se o desafio de lidar com o contexto desbalanceado, além da necessidade da criação de um conjunto de dados anotado focado no domínio tributário, tendo sido empregada uma estratégia de anotação automática de registros. A utilização de Modelos de Linguagem Grandes (do inglês, *Large Language Models*, ou LLMs), baseados em *transformers*, nos experimentos conduzidos destacou a eficácia dessa abordagem na classificação de dados tributários, mesmo diante dos desafios identificados. A partir dos resultados obtidos, observou-se que manter o desbalanceamento no conjunto de dados de treinamento implicou em melhores resultados para o cenário em questão. Além disso, os resultados também indicam que os LLMs com arquitetura *encoder* continuam sendo uma opção eficiente, proporcionando rapidez e compatibilidade com hardware de uso geral. Esses modelos mantêm sua eficácia, mesmo em meio à tendência de preferência por LLMs com arquitetura *decoder* com um número cada vez maior de parâmetros, especialmente em cenários com limitações de recurso de hardware.

Palavras-chave: Processamento de Linguagem Natural, Classificação de Texto, Modelos de Linguagem Grandes, Dados Desbalanceados, Domínio Jurídico, Diários Oficiais.

Abstract

The Legal domain stands as a promising application field for Natural Language Processing. Official Journals contain exceptionally relevant information across various legal subdomains, with significant implications for both public and private sectors. This study used a text classification approach to identify tax-related publications within the Brazilian Official Journal. While analyzing the tax-related context, we addressed the challenge of highly imbalanced data. Our investigation culminated in the creation of an automatically annotated dataset. Using transformer-based Large Language Models (LLMs) in our experiments underscored their suitability for tax-related data classification within the Brazilian Official Journal. Also, our study generated evidence that inserting imbalance into the training set can lead to better results in highly imbalanced contexts. Findings from our study indicate that encoder LLMs remain an efficient choice, offering speed and compatibility with consumer-grade hardware. These models maintain effectiveness even as the prevailing trend leans towards large decoder LLMs.

Keywords: Natural Language Processing, Text Classification, Large Language Models, Imbalanced Data, Legal Domain, Official Journals

Agradecimentos

Agradeço a Deus, por ter me dado forças e me mostrado os caminhos, até mesmo quando foi difícil enxergar.

Agradeço a Fernanda, minha esposa e amor da minha vida, por todo o apoio, paciência e cuidado, sobretudo nos momentos mais difíceis. À você dedico todos os meus feitos, hoje e sempre.

Agradeço a minha família, principalmente meus pais, por sempre terem me ensinado o que é certo e incentivado nas minhas escolhas, quaisquer que fossem. Sou o que sou hoje graças ao que vocês me ensinaram. A vocês, meu amor incondicional.

Ao meu orientador, Professor Cláudio de Souza Baptista, expresso minha eterna gratidão. Palavras jamais serão suficientes para descrever o quanto o senhor já fez por mim. Agradeço por mais essa etapa vencida, sempre com o seu apoio, assim como por todo o incentivo e paciência.

Agradeço ao Professor Luciano Barbosa da UFPE, por ter desprendido do seu tempo para contribuir com este trabalho, sempre com valiosas contribuições. Ao senhor, minha gratidão.

Aos meus amigos e colegas do Laboratório de Sistemas de Informação (LSI), agradeço sinceramente por todo o apoio, aprendizado e parceria. Levarei para sempre os amigos que fiz aqui e os conhecimentos adquiridos.

O meu muito obrigado aos professores do Programa de Pós-Graduação em Ciência da Computação da UFCG, por todos os ensinamentos que levarei para sempre comigo.

Por fim, agradeço à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro parcial à minha formação.

Sumário

Lista de Símbolos

AM - Aprendizagem de Máquina

AUC - Área Sob a Curva, do inglês, *Area Under the Curve*

BoW - *Bag of Words*

DOE - Diário Oficial do Estado

DOM - Diário Oficial do Município

DOU - Diário Oficial da União

FN - Falsos Negativos

FP - Falsos Positivos

IA - Inteligência Artificial

LLM - Modelo de Linguagem Grande, do inglês, *Large Language Models*

NER - Reconhecimento de Entidades Nomeadas, do inglês, *Named Entity Recognition*

NLG - Geração de Linguagem Natural, do inglês, *Natural Language Generation*

NLU - Entendimento de Linguagem Natural, do inglês, *Natural Language Understanding*

PEFT - Ajuste Eficiente de Parâmetros, do Inglês, *Parameter-Efficient Fine-Tuning*

PLN - Processamento de Linguagem Natural

PR - Precisão-Revocação

PR-AUC - Área Sob a Curva de Precisão-Revocação, do inglês, *Precision-Recall Area Under the Curve*

QLoRA - Ajuste Eficiente de LLMs Quantizados, do Inglês, *Efficient Finetuning of Quantized LLMs*

SVM - Máquinas Vetores de Suporte, do inglês, *Support Vector Machines*

TF-IDF - *Term Frequency - Inverse Document Frequency*

VN - Verdadeiros Negativos

VP - Verdadeiros Positivos

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Entidades governamentais utilizam os Diários Oficiais como meio de comunicar decisões, divulgar informações e promover transparência à sociedade, formalizando e tornando públicas ações de relevância e impacto social. Entre os diversos conteúdos publicados, destacam-se leis, resoluções, vetos, portarias, decretos, demonstrações financeiras, editais, licitações e acordos contratuais (MACEDO, 2018; SILVA, 2019). Essa prática estende-se globalmente, como na Europa com o *European Union Official Journal*[1] e o nos Estados Unidos com o *Federal Register*[2]. No contexto brasileiro, o Diário Oficial da União (DOU) assume destaque devido à sua abrangência nacional. Também, cada estado possui seu próprio Diário Oficial do Estado (DOE), bem como alguns municípios contam com seu veículo de divulgação próprio, o Diário Oficial Municipal (DOM).

O conteúdo veiculado em cada diário é categorizado em atos, podendo ser publicado por órgãos da administração pública federal, estadual ou municipal, conselhos profissionais, entidades privadas e, até mesmo, pessoas físicas (PORTARIA..., 2018; QUEM..., 2022). Normalmente, essas publicações são disponibilizadas em websites gerenciados pelos entes governamentais (unidade federativa, estado ou município), em arquivos PDF ou para leitura direta nos próprios sites. A frequência das edições é geralmente diária, contendo um ou mais documentos que podem ultrapassar centenas de páginas, especialmente no caso do DOU. Um fragmento de página extraído diretamente do DOU, composto de diversos atos, pode ser observado na Figura 1.1.

[1] https://eur-lex.europa.eu/homepage.html
[2] https://www.federalregister.gov/

Figura 1.1: Fragmento de página extraído do Diário Oficial da União.

EDUARDO NERY MACHADO FILHO
Diretor-Geral

ACÓRDÃO Nº 57/2023-ANTAQ

1. Processo: 50300.015149/2022-09
2. Interessado: Portos do Paraná - Autoridade Portuária dos Portos de Paranaguá e Antonina
3. Relatora: Flávia Takafashi
4. Unidade Técnica: Superintendência de Regulação - SRG
5. Acórdão:

VISTOS, relatados e discutidos os presentes autos que tratam da revisão tarifária extraordinária do Porto Organizado de Paranaguá/PR,

ACORDAM os Diretores da Agência Nacional de Transportes Aquaviários, reunidos para a Reunião Ordinária de Diretoria Colegiada de nº 537, ante as razões expostas pela Relatora, em:

5.1. homologar e aprovar o pedido da Autoridade Portuária dos Portos de Paranaguá e Antonina para a revisão da estrutura tarifária do Porto de Paranaguá/PR, equivalente a um Índice de Reajuste Médio - IRT de 19,59% e a um Efeito Médio Tarifário - EMT de 24,20%, a qual será efetuada após a ausência de manifestação contrária do Poder Concedente, vencido o período legal de quinze dias úteis após a comunicação desta decisão;

5.2. determinar à Secretaria-Geral - SGE que promova a comunicação junto ao Ministério de Portos e Aeroportos e ao Ministério da Fazenda, nos termos do Ofício-MINUTA GRP (SEI nº 1813655) e do Ofício-MINUTA GRP (SEI nº 1813657), e que publique, após o decurso de quinze dias úteis, a decisão estabelecida nos termos da minuta de Deliberação-DG (SEI nº 1813667);

5.3. determinar à Procuradoria Federal junto à ANTAQ - PFA que adote as medidas necessárias para que seja dado conhecimento do teor da presente decisão ao juízo da 1ª Vara Federal de Francisco Beltrão - Seção Judiciária do Paraná; e

5.4. cientificar a requerente acerca da presente decisão.

6. Data da Reunião: 09/02/2023 - Telepresencial.

7. Especificação do quórum:

7.1. Diretores presentes: Eduardo Nery (Presidente), Flávia Takafashi (Relatora), Lima Filho e Alber Vasconcelos.

EDUARDO NERY MACHADO FILHO
Diretor-Geral

ACÓRDÃO Nº 64/2023-ANTAQ

1. Processo: 50300.017187/2022-98
2. Interessado: Escritório Jurídico Carbone
3. Relator: Eduardo Nery
4. Unidade Técnica: Superintendência de Outorgas - SOG
5. Acórdão:

VISTOS, relatados e discutidos os presentes autos que tratam de consulta acerca do enquadramento de embarcações multipropósito de pesquisa sísmica (Special Purpose Ship - SPS) nas regras estabelecidas pela Agência, em especial no tocante à necessidade de autorização para afretamento,

ACORDAM os Diretores da Agência Nacional de Transportes Aquaviários, reunidos para a Reunião Ordinária da Diretoria Colegiada de nº 537, ante as razões expostas pelo Relator, em:

5.1. informar que as embarcações multipropósito de pesquisa sísmica (Special Purpose Ship - SPS) estão sujeitas à regulação da ANTAQ e às regras e procedimentos estabelecidos na Lei nº 9.432/1997, Resolução Normativa-ANTAQ nº 01/2015 e demais normativos que versam sobre navegação de apoio marítimo; e

5.2. cientificar a interessada acerca da presente decisão.

6. Data da Reunião: 09/02/2023 - Telepresencial.

7. Especificação do quórum:

7.1. Diretores presentes: Eduardo Nery (Presidente e Relator), Flávia Takafashi, Lima Filho, Alber Vasconcelos e Caio Farias.

EDUARDO NERY MACHADO FILHO
Diretor-Geral

SUPERINTENDÊNCIA DE OUTORGAS

DELIBERAÇÃO Nº 42, DE 15 DE FEVEREIRO DE 2023

O SUPERINTENDENTE DE OUTORGAS DA AGÊNCIA NACIONAL DE TRANSPORTES AQUAVIÁRIOS, no uso da competência delegada que lhe é conferida por meio da Portaria DG nº 404-ANTAQ, de 21 de março de 2022, considerando o art. 4º, inciso VII, do Regimento Interno e o que consta do Processo nº 50300.002318/2023-13, resolve:

Art. 1º Declarar extinta, por renúncia, a outorga de titularidade do microempreendedor individual JODEALDO BEZERRIL MENDES 33607745234, inscrito no CNPJ sob o nº 20.898.460/0001-96, constante no Termo de Autorização nº 1.115-ANTAQ, de 2 de fevereiro de 2015.

Art. 2º A extinção da autorização em tela não exime a empresa de eventuais sanções a serem apuradas em regular processo administrativo.

Art. 3º Esta Deliberação-SOG entra em vigor na data de sua publicação.

RENILDO BARROS

Ministério da Previdência Social

SUPERINTENDÊNCIA NACIONAL DE PREVIDÊNCIA COMPLEMENTAR
DIRETORIA DE LICENCIAMENTO

PORTARIA PREVIC Nº 151, DE 10 DE FEVEREIRO DE 2023

O DIRETOR DE LICENCIAMENTO, no uso das atribuições que lhe confere a alínea "d" do inciso I do art. 16 do Decreto nº 11.241, de 18 de outubro de 2022, e considerando as manifestações técnicas exaradas no Processo nº 44011.007358/2022-23, resolve:

Art. 1º Aprovar, com vigência a partir da data de emissão do protocolo pelo sistema informatizado da Previc (licenciamento automático), ocorrida em 14/12/2022, o convênio de adesão celebrado entre a empresa Companhia de Trens Urbanos de Minas Gerais - CBTU-MG, CNPJ nº 46.574.475/0001-92, na condição de patrocinadora do Plano de Contribuição Variável da Patrocinadora CBTU, CNPB nº 2000.0036-56, e a Fundação Rede Ferroviária de Seguridade Social - REFER, CNPJ nº 30.277.685/0001-89, na condição de entidade fechada de previdência complementar responsável pela administração do referido plano.

GEORGE ANDRÉ WILLRICH SALES

PORTARIA PREVIC Nº 152, DE 10 DE FEVEREIRO DE 2023

O DIRETOR DE LICENCIAMENTO, no uso das atribuições que lhe confere a alínea "d" do inciso I do art. 16 do Decreto nº 11.241, de 18 de outubro de 2022, e considerando as manifestações técnicas exaradas no Processo nº 44011.007319/2022-26, resolve:

Art. 1º Aprovar o convênio de adesão celebrado entre o Município de Juiz de Fora - MG, CNPJ nº 18.338.178/0001-02, na condição de patrocinador do Plano Família Previdência Municípios, CNPB nº 2021.0015-47, e a Fundação CEEE de Seguridade Social - ELETROCEEE, CNPJ nº 90.884.412/0001-24, na condição de entidade fechada de previdência complementar responsável pela administração do referido plano.

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

GEORGE ANDRÉ WILLRICH SALES

PORTARIA PREVIC Nº 153, DE 10 DE FEVEREIRO DE 2023

O DIRETOR DE LICENCIAMENTO, no uso das atribuições que lhe confere a alínea "c" do inciso I do art. 16 do Decreto nº 11.241, de 18 de outubro de 2022, e considerando as manifestações técnicas exaradas no Processo nº 44011.008002/2022-15, resolve:

Art. 1º Aprovar as alterações propostas para o estatuto da entidade ENERGISAPREV - FUNDAÇÃO ENERGISA DE PREVIDÊNCIA, CNPJ nº 06.056.449/0001-58, nos termos do supracitado processo.

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

GEORGE ANDRÉ WILLRICH SALES

PORTARIA PREVIC Nº 154, DE 10 DE FEVEREIRO DE 2023

O DIRETOR DE LICENCIAMENTO, no uso das atribuições que lhe confere a alínea "c" do inciso I do art. 16 do Decreto nº 11.241, de 18 de outubro de 2022, e considerando as manifestações técnicas exaradas no Processo nº 44011.005336/2022-29, resolve:

Art. 1º Aprovar as alterações propostas para o estatuto da entidade BEP - CAIXA DE PREVIDÊNCIA SOCIAL - PREVBEP, CNPJ nº 07.697.683/0001-27, nos termos do supracitado processo.

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

GEORGE ANDRÉ WILLRICH SALES

PORTARIA PREVIC Nº 155, DE 10 DE FEVEREIRO DE 2023

O DIRETOR DE LICENCIAMENTO, no uso das atribuições que lhe confere a alínea "d" do inciso I do art. 16 do Decreto nº 11.241, de 18 de outubro de 2022, e considerando as manifestações técnicas exaradas no Processo nº 44011.008177/2022-14, resolve:

Art. 1º Aprovar o 2º termo aditivo ao convênio de adesão celebrado entre a empresa OI S.A., CNPJ nº 76.535.764/0001-43, na condição de patrocinadora do Plano Celprev Amazônia, CNPB nº 2004.0009-29, e a Fundação Atlântico de Seguridade Social - FATL, CNPJ nº 07.110.214/0001-60, na condição de entidade fechada de previdência complementar responsável pela administração do referido plano.

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

GEORGE ANDRÉ WILLRICH SALES

PORTARIA PREVIC Nº 156, DE 10 DE FEVEREIRO DE 2023

O DIRETOR DE LICENCIAMENTO, no uso das atribuições que lhe confere confere a alínea "c" do inciso I do art. 64 da Portaria nº 529, de 8 de dezembro de 2017 (Regimento Interno da Superintendência Nacional de Previdência Complementar - Previc), tendo em conta o disposto no inciso II do art. 25 da Instrução nº 09, de 30 de março de 2022, e considerando as manifestações técnicas exaradas no Processo nº 44011.004248/2021-29, resolve:

Fonte: Diário Oficial da União.

No âmbito dos Diários Oficiais, destaca-se o impacto das publicações de instrumentos normativos, abrangendo leis, decretos, medidas provisórias, dentre outros atos de obrigações legais. Esses instrumentos tornam-se oficiais no momento de sua publicação ou conforme orientações da redação publicada. A integralidade do texto publicado é necessária, contribuindo para a extensão dos documentos. Essas obrigações legais podem envolver diversos ramos do Direito, como tributário, trabalhista, ambiental, dentre outros. Além dos instrumentos normativos, outros subdomínios compreendem o conteúdo dos Diários Oficiais, como

alterações de pessoal, publicações relacionadas a processos licitatórios e contratos públicos, bem como atos e balanços de entidades, tanto públicas como privadas (PORTARIA..., 2018). Destarte, é crucial para os órgãos impactados, sejam públicos ou privados, acompanhar o conteúdo veiculado nos Diários Oficiais, visando sempre a conformidade com a legislação vigente.

Diante desse contexto, destaca-se a importância do monitoramento do conteúdo publicado nos meios oficiais de divulgação, ou seja, os Diários Oficiais. Para realizar esse monitoramento, o primeiro passo consiste na obtenção automática do conteúdo publicado nos Diários Oficiais para análise. No entanto, cada Diário Oficial apresenta desafios específicos, incluindo captchas, preenchimento de formulários para busca, indisponibilidade de PDFs, consulta por página ao invés do documento completo e PDFs não pesquisáveis. Esses desafios foram identificados em uma análise detalhada em 77 Diários Oficiais (Apêndice A), evidenciando o desafio na extração de conteúdo de tais documentos.

Mesmo considerando Diários Oficiais que contêm publicações em formato PDF textual pesquisável, persistem desafios, uma vez que o formato dos documentos não é padronizado (conforme evidenciado no Apêndice B). Além disso, arquivos PDF não fornecem informações sobre o layout, e cada Diário Oficial possui seu próprio formato de disposição do conteúdo nos documentos, com divergências de layout, como fragmentação do conteúdo em colunas, tabelas e diferentes disposições de informações. Essas divergências podem impactar a extração automática do conteúdo de cada ato.

Após a obtenção do conteúdo dos Diários Oficiais, segue-se a inspeção das publicações para identificar o domínio do conteúdo. Os documentos diários, com centenas de páginas, tornam a leitura completa uma atividade custosa, especialmente se realizada por um especialista humano. A busca textual direta também não é uma alternativa viável, pois não há garantia da presença de termos específicos que deveriam ser buscados para abranger todas as publicações de interesse, além da existência de veículos de publicação sem conteúdo pesquisável.

Para além da obtenção e identificação direta do conteúdo publicado em um domínio específico, o tempo de duração desse processo é crucial. Tomemos como exemplo publicações no domínio tributário, especialmente a alteração de alíquotas de impostos. A identificação rápida dessas mudanças, logo após sua publicação, é essencial para o planejamento e execução

da adaptação tempestiva das empresas impactadas. A agilidade na resposta a tais alterações é vital para mitigar o risco de penalidades e interrupções operacionais decorrentes de ajustes nas alíquotas de impostos. A conformidade legal e tributária de empresas diretamente afetadas por mudanças tributárias está intimamente ligada à qualidade do monitoramento dessas alterações em instrumentos regulatórios, todos veiculados nos Diários Oficiais.

Sendo assim, evidencia-se a importância de uma solução capaz de extrair e identificar, tempestivamente, publicações em um domínio específico nos Diários Oficiais. Isso configura um problema passível da aplicação de técnicas de Processamento de Linguagem Natural (PLN). Nesta pesquisa, o contexto de atuação foi o DOU, dada sua grande abrangência e quantidade de publicações. Quanto ao domínio de identificação de publicações, optou-se pelo domínio tributário, dada sua relevância e frequentes alterações, como também pela existência de um conjunto de dados anotado com atos tributários.

A análise do conteúdo do DOU revela-se uma atividade de extrema importância para diversas empresas que buscam garantir sua conformidade tributária, especialmente aquelas com instalações em vários locais do país, como grandes varejistas. No entanto, a análise automatizada do DOU é complexa, como exposto anteriormente, somada ao alto volume de publicações. Destaca-se também o desbalanceamento na frequência de publicações tributárias, o que representa um desafio adicional, pois apenas uma fração mínima das inúmeras publicações diárias contém conteúdo relacionado ao domínio tributário.

O Diário Oficial da União possui uma frequência diária de publicações na casa dos milhares, portanto, faz-se necessário que o fluxo de processamento tenha o baixo tempo de execução como um de seus requisitos. Apesar de haver intervenção humana, essa abordagem ainda assim reduz a carga de trabalho da análise de centenas de milhares de publicações para apenas algumas centenas. Na Tabela 1.1 são apresentados alguns exemplos de publicações tributárias e não tributárias extraídas do Diário Oficial da União.

A importância desse problema transcende as questões operacionais imediatas das empresas, com implicações mais amplas no cenário econômico e jurídico. A eficiência na gestão tributária é vital para o ambiente de negócios, influenciando diretamente a competitividade, a sustentabilidade financeira e a conformidade legal das organizações. A identificação ágil e precisa de alterações nas leis tributárias, veiculadas no DOU, não apenas protege as empresas de consequências adversas, mas também contribui para a construção de um ambiente de

Tabela 1.1: Fragmentos extraídos de exemplos de publicações veiculadas no Diário Oficial da União.

Fragmento do Texto Publicado	Classificação
"[...] Altera o Ato COTEPE/ICMS n° 3/22, que divulga relação de produtores de B100 optantes pelo tratamento tributário diferenciado para apuração e pagamento do ICMS incidente nas operações com B100 realizadas com diferimento ou suspensão, na forma do Convênio ICMS nº 206/21. [...]"	Tributário
"[...] Dispõe sobre a emissão de Letra de Risco de Seguro (LRS) por Sociedade Seguradora de Propósito Específico (SSPE), sobre as regras gerais aplicáveis à securitização de direitos creditórios e à emissão de Certificados de Recebíveis e sobre a flexibilização do requisito de instituição financeira para a prestação do serviço de escrituração e de custódia de valores mobiliários: [...]"	Não tributário
"[...] Altera para zero por cento as alíquotas do Imposto de Importação incidentes sobre os Bens de Capital que menciona, na condição de Ex-tarifários. [...]"	Tributário
"[...] O DIRETOR DO DEPARTAMENTO NACIONAL DE REGISTRO EMPRESARIAL E INTEGRAÇÃO, no uso das atribuições que lhe confere o art. 4º, da Lei nº 8.934, de 18 de novembro de 1994, resolve: Art. 1º A Instrução Normativa DREI nº 82, de 2021, passa a vigorar com as seguintes alterações: Art. 3º Os livros de que trata o art. 1º deverão ser exclusivamente digitais, podendo ser produzidos ou lançados em plataformas eletrônicas. § 1º Os sistemas eletrônicos utilizados devem garantir, no mínimo, a segurança, a confiabilidade e a inviolabilidade dos dados. [...]"	Não tributário
"[...] Art. 1º A alocação da cota para importação estabelecida pela Resolução do Comitê-Executivo de Gestão da Câmara de Comércio Exterior nº 400, de 22 de setembro de 2022, publicada no Diário Oficial da União (DOU) de 23 de setembro de 2022, consignada no Anexo Único desta Portaria, será realizada em conformidade com as seguintes regras: [...]"	Tributário
"[...] Esta Instrução Normativa institui o Manual de Cobrança, Recuperação e Parcelamento de Créditos do CNPq, que regulamenta os procedimentos para cobrança, recuperação e parcelamento de créditos, tanto para pessoas físicas, ex-beneficiárias de auxílios financeiros à pesquisa ou de bolsas no país e/ou no exterior, quanto para pessoas jurídicas, inadimplentes com o CNPq. [...]"	Não tributário
"[...] Altera a Lei nº 9.478, de 6 de agosto de 1997, e a Lei nº 9.718, de 27 de novembro de 1998, para promover ajustes na cobrança da Contribuição para os Programas de Integração Social e de Formação do Patrimônio do Servidor Público - PIS/Pasep e da Contribuição para o Financiamento da Seguridade Social - Cofins incidentes sobre a cadeia de produção e de comercialização de etanol hidratado combustível. [...]"	Tributário
"[...] A escritura particular pode ser feita e assinada ou somente assinada pelos contratantes, sendo subscrita por 2 (duas) testemunhas, observado que as assinaturas poderão ser feitas de forma eletrônica, conforme legislação aplicável. [...] Após a apresentação da contestação pelo expropriado, se não houver oposição expressa com relação à validade do decreto desapropriatório, deverá ser determinada a imediata transferência da propriedade do imóvel para o expropriante, independentemente de anuência expressa do expropriado, e prosseguirá o processo somente para resolução das questões litigiosas. [...]"	Não tributário

Fonte: De autoria própria, a partir de dados extraídos do DOU.

negócios com mais *compliance* (VITALIS, 2019; GUERRA; GUERRA, 2023).

Portanto, a abordagem estratégica para lidar com a complexidade da análise do DOU, principalmente no contexto tributário, não só atende às necessidades operacionais imediatas, mas também aponta para um entendimento mais profundo da interseção entre regulamentação fiscal e práticas comerciais. Esse entendimento é crucial para sustentar a integridade e a conformidade das operações empresariais em um cenário dinâmico e sujeito a mudanças constantes.

1.1 Objetivos

O objetivo geral deste trabalho é de prover uma solução automatizada para a obtenção e classificação de atos publicados no DOU referentes ao domínio tributário. Para atingir esse objetivo geral, o trabalho foi desdobrado em objetivos específicos, a saber:

1. Elaborar uma solução automatizada para a obtenção de dados do DOU, visando a construção de um *corpus* de propósito geral, o monitoramento de frequência diária e o armazenamento de atos para posterior classificação;

2. Criar um conjunto de dados com base no DOU, direcionado para a atividade de classificação automática no ramo tributário, a partir do *corpus* do DOU obtido no objetivo anterior; e

3. Treinar e avaliar modelos, utilizando técnicas de Aprendizagem de Máquina e Processamento de Linguagem Natural, com ênfase na tarefa de classificação de atos do Diário Oficial da União no domínio tributário.

1.2 Questões de Pesquisa

Com o andamento da pesquisa, deseja-se responder as seguintes questões de pesquisa:

1. Os Modelos de Linguagem Grandes (do inglês, *Large Language Models*, ou LLMs) são adequados para a tarefa de classificação de atos de domínio tributário no DOU?

2. Quão bons são os LLMs pré-treinados de domínio geral quando comparados com LLMs pré-treinados em *corpora* do domínio do problema (textos jurídicos e, mais especificamente, documentos do DOU) para a atividade de classificação de texto tributário no DOU?

3. Quais os impactos das diferentes proporções de balanceamento nos dados de treinamento no contexto desbalanceado da classificação de publicações tributárias no DOU?

4. Frente ao grande foco recente nos LLMs de arquitetura *decoder*, os LLMs *encoder* apresentam resultados competitivos, mesmo possuindo menos parâmetros, quando comparados a modelos *decoder* da ordem de 7 bilhões de parâmetros?

1.3 Contribuições

Com base na pesquisa realizada neste estudo, é fundamental destacar as diversas contribuições obtidas. Inicialmente, foram empregadas técnicas de PLN de estado da arte em um contexto inovador, que envolve a classificação de texto extraído do DOU no domínio tributário. Adicionalmente, ao abordar um contexto de dados desbalanceados, foram geradas evidências do impacto de diferentes proporções aplicadas a um conjunto de treinamento de classificação binária. Contextos com alto desbalanceamento são comuns ao lidar com cenários de aplicação reais, especialmente na área jurídica, que é o domínio da linguagem empregada nos Diários Oficiais. Os experimentos realizados neste trabalho indicam que introduzir desbalanceamento em conjuntos de treinamento pode influenciar positivamente nos resultados.

Outra contribuição é a comparação de resultados no domínio do direito tributário entre LLMs pré-treinados em domínios específicos, como o jurídico ou do DOU, e aqueles pré-treinados em um domínio geral. No problema abordado, LLMs pré-treinados em domínios alinhados com a tarefa-alvo não resultaram em melhorias significativas. Essa constatação sustenta a conclusão de que existem cenários nos quais o domínio de pré-treinamento do modelo não é determinante nos resultados, especialmente quando outros fatores, como léxico específico da linguagem do domínio, como no caso da linguagem jurídica e do DOU, e alto desbalanceamento, exercem influência.

Por fim, os resultados deste trabalho sustentam as conclusões de que LLMs de arquitetura *encoder*, mesmo tendo uma menor quantidade de parâmetros frente aos LLMs *decoder* da ordem de 7 bilhões de parâmetros, ainda consistem em alternativas competitivas. Modelos transformer de arquitetura *decoder*, justamente pela alta quantidade de parâmetros, frequentemente estão associados a um alto consumo de recursos computacionais. Sendo assim, torna-se um desafio a execução de experimentos, bem como a sua aplicação em cenários reais, sem a utilização de alternativas para torná-los mais eficientes, muitas vezes em detrimento da qualidade da representação. Modelos *encoder* caracterizaram uma alternativa mais eficiente e capaz de obter melhores resultados no contexto da classificação binária altamente desbalanceada no domínio tributário em meio ao DOU.

1.4 Estrutura da Dissertação

O restante desta dissertação está estruturado conforme descrito a seguir.

No Capítulo 2, contemplam-se os principais conceitos teóricos e soluções para a compreensão do contexto da pesquisa. Este capítulo constitui a base para a discussão dos resultados e interpretação das descobertas obtidas ao longo da pesquisa.

O Capítulo 3 contém a exploração de pesquisas e abordagens anteriores relacionadas ao tema desta dissertação. A revisão da literatura fornece um panorama das contribuições existentes, destacando lacunas no conhecimento que motivaram a presente investigação. Este capítulo é fundamental para situar a pesquisa no contexto mais amplo do campo de estudo.

O Capítulo 4 contém o detalhamento da abordagem metodológica adotada para realizar a pesquisa. São discutidos os materiais e métodos utilizados na coleta, modelagem e análise dos dados e resultados obtidos.

No Capítulo 5, está contida a análise e discussão dos resultados obtidos. Este capítulo visa destacar as descobertas da pesquisa, analisando criticamente os dados em relação aos objetivos propostos. São exploradas as nuances dos experimentos realizados, contextualizando os resultados dentro do arcabouço teórico apresentado anteriormente.

Finalmente, no Capítulo 6, estão apresentadas as principais conclusões derivadas da pesquisa, ressaltando as limitações, respostas às questões de pesquisa e possíveis direções para futuras pesquisas. Este capítulo encerra a dissertação, proporcionando uma síntese do traba-

lho realizado, além de consolidar os principais achados em um contexto mais amplo.

Capítulo 2

Fundamentação

Neste capítulo é apresentada a base teórica que sustenta a compreensão e análise da temática deste trabalho. O conteúdo está organizado da seguinte forma: na Seção 2.1 é apresentado o conceito de Direito Tributário; na Seção 2.2 são detalhadas as características do Diário Oficial da União; na Seção 2.3 aborda-se o Processamento de Linguagem Natural (PLN); na Seção 2.4 são apresentados modelos de classificação tradicionais aplicados ao PLN; na Seção 2.5 é discorrido acerca dos Modelos de Linguagem Grandes; tratando-se da Seção 2.6, são discutidas as métricas de avaliação para soluções de Aprendizagem de Máquina (AM); por fim, a Seção 2.7 é objeto das considerações finais deste capítulo.

2.1 Direito Tributário

O direito tributário consiste no "ramo do direito que rege as relações jurídicas entre o Estado e os particulares, decorrentes da atividade financeira do Estado no que se refere à obtenção de receitas que correspondam ao conceito de tributos" (BECHO, 2017). Diversas normas jurídicas regulam a tributação incidente nas mais variadas esferas e relações interpessoais, sendo estas físicas, jurídicas ou o próprio Estado. Sendo assim, o direito tributário tem como objeto de estudo o complexo cenário de regras de arrecadação, instituição e fiscalização de tributos (FOLLONI; SIMM, 2016).

A complexidade atrelada a esse ramo do direito não se limita apenas ao vasto conjunto de dispositivos e normas jurídicas regendo o contexto tributário, especialmente o brasileiro, mas estende-se também nos impactos sociais decorrentes da sua aplicação. Dada sua importância

na normatização da arrecadação financeira do Estado, todos os componentes da sociedade, independentemente de classe social, são impactados pela legislação tributária vigente. Esse fato caracteriza a interdisciplinaridade atrelada ao Direito Tributário, com altos impactos sociais, caminhando em conjunto com áreas como a Economia e a Contabilidade (BECHO, 2017).

O Código Tributário Nacional (CTN), publicado na Lei Nº 5.172, de 25 de outubro de 1966 (LEI..., 1966) e em conjunto com suas subsequentes alterações ao longo dos anos, regula o sistema tributário nacional. O CTN também contém as normas gerais de direito tributário passíveis de aplicação em todo o território nacional, tanto pelas instituições da União, Estados, Municípios e Distrito Federal.

O cenário tributário brasileiro é alvo de análises de relatórios de respeitados bancos internacionais, como o Banco Mundial[1]. Um de seus relatórios periodicamente publicados é o *Doing Business*, que realiza um comparativo entre 190 países no que diz respeito a instrumentos regulatórios, dentre eles o tributário. O relatório busca avaliar os países de forma a identificar os que possuem as melhores condições de negócios, estabelecendo uma classificação para cada país. Para o relatório do ano de 2020, último ano de veiculação disponível e que contém dados de 2019, o Brasil ocupava a 124ª posição, logo abaixo do Senegal (PORTO, 2019). Ou seja, a complexidade regulatória brasileira, a 9ª economia do mundo em termos de Produto Interno Bruto no período da pesquisa segundo dados do próprio Banco Mundial[2], encontra-se comparável a do Senegal, 111ª economia do mundo no mesmo período.

Outro importante relatório do Banco Mundial, o *Paying Taxes*, compara os diferentes sistemas tributários ao redor do mundo (PAYING..., 2018; PAYING..., 2020). Um importante indicador de complexidade apontado pelo relatório é o *time to comply*, ou tempo de conformidade, que representa o tempo investido pelas empresas de forma a realizar suas atividades tributárias em busca de conformidade fiscal, como as obrigações acessórias. Entre os anos de 2004 e 2018 observou-se uma redução de 23% em termos de *time to comply* no Brasil, onde grande parte desta evolução atribui-se à introdução e aprimoramento de sistemas eletrônicos (PAYING..., 2020). Ainda assim, o índice brasileiro para o ano de 2018 era 8,2 vezes maior do que a média mundial (PAYING..., 2018), configurando-se como um exemplo negativo,

[1] https://www.worldbank.org/en/home
[2] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

com uma alta complexidade tributária e alto tempo de conformidade quando comparado ao cenário mundial (PORTO, 2019).

Diante desse cenário, observa-se a importância do direito tributário na sociedade brasileira. Além de haver margem para melhorias, lidar com essa complexidade é necessário e essencial para que todos os cumprimentos legais sejam atendidos, tanto por cidadãos comuns como por empresas. Inovações tecnológicas e ferramentas que auxiliem no processo de atendimento às obrigações fiscais são essenciais para a manutenção do processo, bem como a evolução e crescimento do país como um todo.

2.2 Diário Oficial da União

Originado ainda no Brasil Imperial, o DOU[3], então chamado Diário Official do Império do Brasil, em 1862, configura-se como o periódico oficial de publicização de atos governamentais de escopo nacional (MENEZES; A, 2019). O DOU foi veiculado de forma impressa por 155 anos, tendo uma versão digital a partir de 2013 e tornando-se inteiramente digital em 2017 (MENEZES; A, 2019). Atualmente, o DOU é gerenciado pelo órgão federal denominado Imprensa Nacional, que, dentre suas responsabilidades, possui o objetivo de publicar, preservar e divulgar os atos oficiais da administração pública federal. A Figura 2.1 contém um exemplo de página inicial do DOU.

A publicação de conteúdo no DOU é realizada por órgãos públicos e entidades privadas, conforme disposto na Portaria IN/SG/PR Nº 9, de 4 de fevereiro de 2021 (PORTARIA..., 2021). Essa portaria também aponta as normas de publicação no DOU, sua subdivisão em três cadernos, ou seções, a temática de cada caderno, e quais os tipos de atos respectivos a cada um dos cadernos. Cada seção possui seu próprio arquivo isolado de publicação, estando os atos tributários inseridos no conteúdo da seção 1. Segue abaixo, conforme publicado na portaria referenciada, o conteúdo respectivo de cada seção:

1. Atos da Seção 1:

    (a) decisões relativas ao controle de constitucionalidade pelo Supremo Tribunal Federal;

[3]https://www.gov.br/imprensanacional/pt-br/acesso-a-informacao/institucional/competencias

Figura 2.1: Exemplo de página inicial do DOU, extraída da edição de 03/01/2024.

ISSN 1677-7042

DIÁRIO OFICIAL DA UNIÃO

REPÚBLICA FEDERATIVA DO BRASIL • IMPRENSA NACIONAL

Ano CLXII Nº 2 — Brasília - DF, quarta-feira, 3 de janeiro de 2024 — SEÇÃO 1

Sumário

Presidência da República

CASA CIVIL

INSTITUTO NACIONAL DE TECNOLOGIA DA INFORMAÇÃO

DESPACHOS

DEFIRO o credenciamento da MAP AUTORIDADE DE REGISTRO E CERTIFICAÇÃO DIGITAL. Processo nº 00100.003279/2023-50.

DEFIRO, a pedido, o descredenciamento da AR LS SOLUTION CERTIFICADO DIGITAL. Processo nº 00100.003272/2023-38.

INDEFIRO o credenciamento da AC DIGITAL MÚLTIPLA G2. Processo nº 00100.001953/2023-61.

INDEFIRO o credenciamento da AR QUALITYCERT DIGITAL CERTIFICADOS LTDA. Processo nº 00100.002213/2023-42.

DEFIRO o descredenciamento da AR HIGH TECH SOLUÇÕES DIGITAIS. Processo nº 00100.003067/2023-72.

MAURÍCIO AUGUSTO COELHO
Diretor-Presidente
Substituto

IMPRENSA NACIONAL

PORTARIA IN/CC/PR Nº 1, DE 2 DE JANEIRO DE 2024

Dispõe sobre os procedimentos de cadastramento, pagamento e publicação de atos no Diário Oficial da União, e dá outras providências.

O DIRETOR-GERAL DA IMPRENSA NACIONAL, no uso das atribuições que lhe conferem os arts. 16 e 20 do Decreto nº 9.215, de 29 de novembro de 2017, resolve:

CAPÍTULO I
DO OBJETO E DO ÂMBITO DE APLICAÇÃO

Art. 1º Esta Portaria disciplina os procedimentos de cadastramento, pagamento e publicação de atos no Diário Oficial da União.

Parágrafo único. Para os fins desta Portaria, são adotadas as seguintes definições:

I - origem - o órgão ou a entidade, a pessoa jurídica ou a pessoa natural, emitente ou demandante do ato a ser publicado no Diário Oficial da União;

II - gerente INCom - o operador que, representando um órgão, entidade ou particular é responsável pelo envio de atos e pela solicitação de cadastramento de Usuários, vinculado a uma Origem;

III - usuário - o operador cadastrado pela Imprensa Nacional para o envio de atos, estando vinculado a uma Origem; e

IV - cliente - denominação genérica para todos os que utilizam os serviços prestados pela Imprensa Nacional.

CAPÍTULO II
DO CADASTRAMENTO

Art. 2º Compete à Imprensa Nacional proceder ao cadastramento:

I - da origem emitente ou demandante do ato a ser publicado no Diário Oficial da União;

II - de Gerente INCom dos interessados em publicar no Diário Oficial da União; e

III - de Usuários.

Parágrafo único. O cadastramento de Origens, de Gerentes INCom e de Usuários dar-se-á por meio do Sistema Eletrônico de Envio de Matérias - INCom da Imprensa Nacional.

Art. 3º Terão suas Origens cadastradas automaticamente por meio do Sistema de Organização e Inovação Institucional do Governo Federal - SIORG:

I - os órgãos da União, independente do Poder que integrarem;

II - as seguintes entidades:

a) as autarquias federais;

b) as fundações públicas federais.

c) as fundações federais de direito privado; e

d) as empresas estatais dependentes de recursos do Tesouro Nacional para o custeio de despesas de pessoal ou para o custeio em geral.

Parágrafo único. Para fins de cadastramento de Gerente INCom e de Usuário dos órgãos e entidades mencionados no **caput**, é necessário o registro, em sistema específico da Imprensa Nacional, dos seguintes documentos:

I - ofício de solicitação para cadastramento, automaticamente gerado pelo sistema, e assinado digitalmente por meio da Plataforma GOV.BR pelo representante legal da instituição e pelo Gerente INCom a ser cadastrado;

II - ato de nomeação, designação ou similar, em que conste o cargo, posto ou emprego do servidor, do militar ou do empregado que está representando o órgão ou a entidade; e

III - ficha de atualização, gerada automaticamente pelo sistema, contendo os dados atualizados da entidade e do Gerente INCom a ser cadastrado.

Art. 4º Depende de solicitação o cadastramento das seguintes Origens:

I - empresas estatais não dependentes dos recursos do Tesouro Nacional para o custeio de despesas de pessoal ou para o custeio em geral;

II - órgãos e entidades de outros entes federados;

III - pessoas jurídicas de direito público externo;

IV - pessoas jurídicas de direito privado;

V - conselhos profissionais;

VI - serviços sociais autônomos; e

VII - pessoas naturais.

Art. 5º Para fins de cadastramento de Origem das instituições mencionadas nos incisos I, II e III do art. 4º, é necessária a apresentação dos seguintes documentos:

I - ofício de solicitação para cadastramento, automaticamente gerado pelo sistema e assinado digitalmente por meio da Plataforma GOV.BR pelo representante legal da instituição requerente;

II - ato de nomeação, designação ou similar, em que conste o cargo do servidor, o posto do militar ou o emprego do trabalhador que representa legitimamente a instituição solicitante;

III - comprovante de pagamento da tarifa de cadastramento, salvo se no caso do inciso III do art. 4º; e

IV - certidão negativa de débitos anteriores, emitido pela Coordenação de Orçamento e Finanças, da Coordenação-Geral de Administração da Imprensa Nacional, salvo se no caso do inciso III do art. 4º.

Parágrafo único. Para fins de cadastramento de Gerente INCom e de Usuário das instituições mencionadas no **caput**, são necessários os seguintes documentos:

I - ofício de solicitação para cadastramento, automaticamente gerado pelo sistema e assinado digitalmente por meio da Plataforma GOV.BR pelo representante legal da instituição, ou procurador, e pelo Gerente INCom a ser cadastrado;

II - ato de nomeação, designação ou similar, em que conste o cargo do servidor, o emprego do trabalho ou o posto do militar que está representando legalmente a instituição, ou mandato, nos termos da lei;

III - ficha de atualização, gerada automaticamente pelo sistema, contendo os dados atualizados da instituição e do Gerente INCom a ser cadastrado; e

IV - pagamento da tarifa de cadastramento, salvo se no caso do inciso III do art. 4º.

Art. 6º Para fins de cadastramento de Origem das instituições indicadas nos incisos IV, V e VI do **caput** do art. 4º desta Portaria, é necessária a apresentação dos seguintes documentos:

I - ofício de solicitação para cadastramento, automaticamente gerado pelo sistema e assinado digitalmente por meio da Plataforma GOV.BR e pelo representante legal da instituição requerente, ou procurador;

II - contrato social ou estatuto acompanhado da Ata de posse da diretoria vigente, em que constem os dados do representante legal que assina o ofício citado no item anterior, ou mandato, nos termos da lei;

III - comprovante de pagamento da taxa de cadastramento; e

IV - certidão negativa de débitos anteriores, emitido pela Coordenação de Orçamento e Finanças, da Coordenação-Geral de Administração da Imprensa Nacional.

Parágrafo único. Para fins de cadastramento de Gerente INCom e de Usuário das instituições mencionadas no **caput**, é necessária a apresentação dos seguintes documentos:

I - ofício de solicitação para cadastramento, automaticamente gerado pelo sistema e assinado digitalmente por meio da Plataforma GOV.BR pelo representante legal da instituição, ou procurador, e pelo Gerente INCom a ser cadastrado;

II - contrato social ou estatuto acompanhado da Ata de posse da diretoria vigente, em que conste os dados do representante legal, que assina o ofício citado no item anterior, ou mandato, nos termos da lei;

III - ficha de atualização, gerada automaticamente pelo sistema, em que conste os dados atualizados da instituição e do Gerente INCom a ser cadastrado; e

IV - pagamento da tarifa de cadastramento.

Art. 7º Para fins de cadastramento de pessoas naturais, é necessária a apresentação dos seguintes documentos:

I - ofício de solicitação de cadastramento, automaticamente gerado pelo sistema e assinado digitalmente por meio da Plataforma GOV.BR;

II - comprovante de pagamento da tarifa de cadastramento; e

III - certidão negativa de débitos anteriores, emitido pela Coordenação de Orçamento e Finanças, da Coordenação-Geral de Administração da Imprensa Nacional.

§ 1º No caso das pessoas naturais, o cadastramento será único para as funções de Origem, Gerente INCom e Usuário.

§ 2º Os atos para publicação no Diário Oficial da União oriundos de pessoas naturais estão restritos àqueles de natureza particular, consideradas as vedações dispostas no art. 35 desta Portaria.

Art. 8º Será emitido pela Imprensa Nacional, após a efetivação do cadastramento, certificado digital individual para cada Gerente INCom.

Parágrafo único. Em caso de interoperabilidade entre sistemas, nos termos do § 1º do art. 36, deverá ser emitido pela Imprensa Nacional certificado digital para o equipamento a ser utilizado na transmissão de atos.

Art. 9º O certificado de que trata o art. 8º deverá obedecer preferencialmente ao padrão Infraestrutura de Chaves Públicas Brasileira ICP-Brasil.

Art. 10. As contas cadastradas somente serão ativadas após a emissão do certificado de que trata o **caput** do art. 8º desta Portaria.

Art. 11. O certificado digital terá validade de cinco anos.

Art. 12. A Imprensa Nacional procederá à atualização da base cadastral a cada cinco anos, a partir da entrada em vigor desta Portaria, por meio dos seguintes órgãos componentes de sua estrutura:

I - Coordenação-Geral de Publicação, Produção e Preservação; e

II - Coordenação-Geral de Tecnologia da Informação.

§ 1º Os procedimentos para a atualização cadastral das Origens e dos Gerentes INCom serão disponibilizados por meio do Portal da Imprensa Nacional.

§ 2º Para fins de atualização, serão cobrados os serviços de cadastramento, na forma do § 2º do art. 17 do Decreto nº 9.215, de 2017.

Art. 13. Será permitida a alteração de registro entre Origens para Gerentes INCom já cadastrados que mudem de local de trabalho apenas quando se processarem entre as seguintes instituições:

I - órgãos da União, independente do Poder que integrarem;

II - autarquias federais;

III - fundações públicas federais; ou

IV - empresas estatais dependentes de recursos do Tesouro Nacional para o custeio de despesas de pessoal ou para o custeio em geral.

Parágrafo único. A solicitação de mudança de Origem para Gerente INCom já cadastrado deverá ser feita pelo representante legal da nova lotação, mediante envio dos documentos comprobatórios constantes do parágrafo único do art. 5º.

Art. 14. Somente os Gerentes INCom cadastrados junto à Imprensa Nacional poderão enviar atos para fins de publicação.

Art. 15. As pessoas jurídicas interessadas em atuar na intermediação para transmissão de atos junto ao sistema da Imprensa Nacional deverão solicitar seu cadastramento e apresentar os documentos indicados nos incisos do **caput** do art. 6º desta Portaria.

Parágrafo único. As pessoas jurídicas intermediadas deverão conceder autorização, válida por até cinco anos, mediante formulário próprio disponibilizado por sistema informatizado da Imprensa Nacional.

Fonte: Diário Oficial da União

(b) os atos com conteúdo normativo da União, das autarquias, das fundações públicas, das empresas públicas e das sociedades de economia mista, exceto os atos de aplicação exclusivamente interna que não afetem interesses de terceiros;

(c) os pareceres do Advogado-Geral da União de que trata o art. 40, § 1º, da Lei Complementar nº 73, de 10 de fevereiro de 1993;

(d) atos do Tribunal de Contas da União, de interesse geral;

(e) atos normativos do Poder Judiciário, do Ministério Público da União e da Defensoria Pública da União, excetuando-se os de caráter interno; e

(f) atas dos órgãos dos Poderes da União com publicidade exigida por legislação específica.

2. Atos da Seção 2:

(a) atos relativos a pessoal da União, das autarquias, das fundações públicas, das empresas públicas e das sociedades de economia mista, cuja publicação decorra de disposição legal.

3. Atos da Seção 3:

(a) decisões relativas ao controle de constitucionalidade pelo Supremo Tribunal Federal;

(b) extratos de instrumentos contratuais e congêneres, de convênios, de dispensa e de inexigibilidade de licitação, de distrato, de registro de preços, de rescisão;

(c) os editais de citação, de intimação, de notificação e de concursos públicos;

(d) os comunicados, os avisos de licitação, de dispensa e de inexigibilidade de licitação, de registro de preços, de anulação, de revogação e os resultados

(e) de julgamentos, entre outros atos da administração pública, cuja publicação seja exigida por determinação legal ou decorrente de norma infralegal; e

(f) atos de pessoas jurídicas de direito privado em geral e de pessoas físicas que tenham como objetivo atender às exigências de publicidade constantes da legislação.

2.3 Processamento de Linguagem Natural

O PLN, na Ciência da Computação, consiste em uma área de pesquisa direcionada a conceber soluções, sistemas e algoritmos capazes de interagir com a linguagem humana (LAURIOLA; LAVELLI; AIOLLI, 2022). Além de ter como foco fazer com que computadores entendam a linguagem natural, o PLN também objetiva utilizar-se do poder computacional para otimizar processos e o trabalho realizado por humanos envolvendo a linguagem natural. Essa área de pesquisa pode ser dividida em duas principais partes: entendimento de linguagem natural e geração de linguagem natural (ou, do inglês, *Natural Language Understanding* e *Natural Language Generation*, respectivamente) (KHURANA *et al.*, 2022a).

Dentre algumas aplicações de PLN, é possível ressaltar desde tradução de texto e detecção de spam em e-mails até extração de informações, geração e sumarização de texto. Outras aplicações relevantes são: classificação de texto, chatbots, reconhecimento de entidades nomeadas, resposta a perguntas e análise de sentimentos. Além disso, diversos domínios são alvo de soluções de PLN, contribuindo ativamente para melhoria de qualidade de serviços, bem como trazendo eficiência para atividades atreladas ao uso de linguagem natural. Dentre os domínios frequentes de atuação de soluções de PLN, compreende-se: medicina, educação, agricultura e internet das coisas, por exemplo (PRIYA; NANDHINI; GNANASEKARAN, 2021; KHURANA *et al.*, 2022a; MAH; SKALNA; MUZAM, 2022).

Entretanto, as aplicações não limitam-se aos domínios citados, dada a vastidão de atividades humanas que utilizam-se de linguagem natural. Com o passar do tempo, em conjunto com o advento e a evolução de novas tecnologias, técnicas de Aprendizagem de Máquina (AM), Inteligência Artificial e Aprendizagem Profunda passaram a ser incorporadas a soluções de PLN, aprimorando e trazendo resultados cada vez mais disruptivos frente ao estado da arte (LAURIOLA; LAVELLI; AIOLLI, 2022; KHURANA *et al.*, 2022a). Obtém-se, pois, soluções capazes de revolucionar a forma de utilização do PLN no dia a dia, a exemplo do ChatGPT[4] e ferramentas similares, como o Bard[5], que refletiram em impactos diretos nos cidadãos comuns, não limitando-se a academia ou indústria (DWIVEDI *et al.*, 2023).

[4]https://chat.openai.com/
[5]https://bard.google.com/

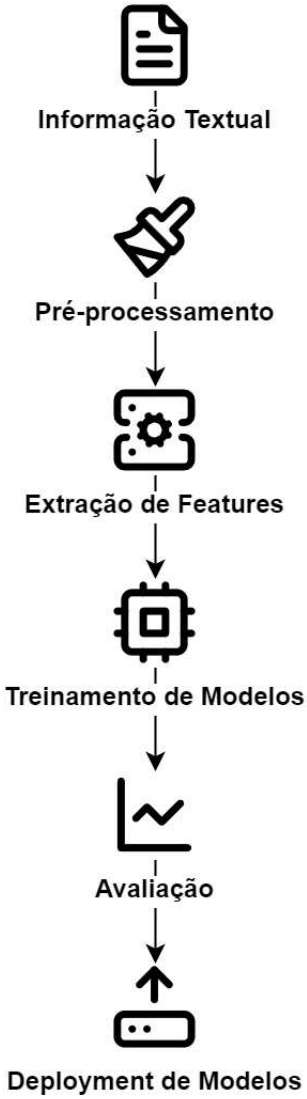
2.4 Modelos de Classificação Tradicionais Aplicados ao PLN

Diversas técnicas e ferramentas de PLN surgiram ao longo dos anos com o objetivo de resolver problemas relacionados à interpretação, análise e manipulação da linguagem natural. Redes neurais passaram a ser amplamente utilizadas e associadas à resolução de problemas de PLN por volta do ano de 2013 (KHURANA *et al.*, 2022b), o que, ao longo dos anos seguintes, levou ao desenvolvimento de técnicas avançadas, como os modelos *transformers*. No entanto, técnicas tradicionais têm sido aplicadas em problemas de modelagem da linguagem natural há décadas, como é o caso de modelos como Naïve Bayes, Árvores de decisão e Support Vector Machines (SVMs) (KHURANA *et al.*, 2022b).

A utilização de técnicas, algoritmos e modelos de AM em geral, aplicados a problemas de linguagem natural, constitui a forma tradicional de resolução desta classe de problemas. No entanto, essa abordagem está associada a etapas adicionais focadas em processar dados do formato de linguagem natural para representações vetoriais, por exemplo, extraindo características e fornecendo recursos para utilização em modelos de classificação (KHURANA *et al.*, 2022b). A Figura 2.2 apresenta a descrição de um fluxo de processamento de soluções de PLN tradicionais, englobando cinco etapas: pré-processamento, extração de características, treinamento de modelos, avaliação e *deployment* (FERRARIO; NAEGELIN, 2020; KHURANA *et al.*, 2022b).

A etapa inicial, de pré-processamento, engloba a limpeza e, como o próprio nome já indica, o pré-processamento do texto, incluindo a remoção de caracteres desnecessários, a conversão para letras minúsculas, a tokenização e a remoção de palavras irrelevantes (*stopwords*). Na etapa de extração de características, o objetivo é extrair características relevantes do texto pré-processado, como as técnicas de *Bag-of-Words* como TF-IDF (*Term Frequency-Inverse Document Frequency*) e *Word Embeddings*. Na etapa de treinamento de modelos, podem ser utilizados diferentes algoritmos e técnicas de classificação, como Naïve Bayes, SVM e XGBoost. Quanto à avaliação dos modelos, são necessárias métricas adequadas para avaliar o desempenho dos modelos em relação ao treinamento aplicado, aos hiperparâmetros e aos dados utilizados. A etapa final, o *deployment*, consiste na utilização do modelo em cenários de aplicações reais, processando dados novos e nunca antes vistos

Figura 2.2: Fluxo de execução de soluções de PLN.



Fonte: De autoria própria.

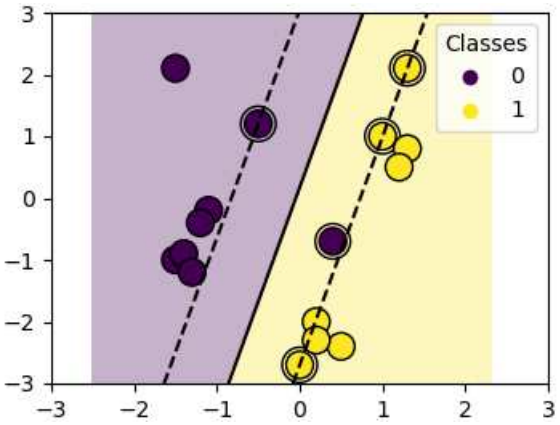
pelo modelo.

2.4.1 *Support Vector Machines*

O SVM é um modelo de AM que pode ser utilizado tanto para tarefas de classificação quanto de regressão. Provou-se ser um método versátil, sendo amplamente utilizado na academia (CERVANTES *et al.*, 2020; KAMRAN; SAEED; ALMAGHTHAWI, 2023; ABDALLA; AMER; RAVANA, 2023). O objetivo do SVM consiste em separar diferentes classes utilizando uma estratégia baseada em hiperplanos, definindo uma superfície capaz de separar e, portanto,

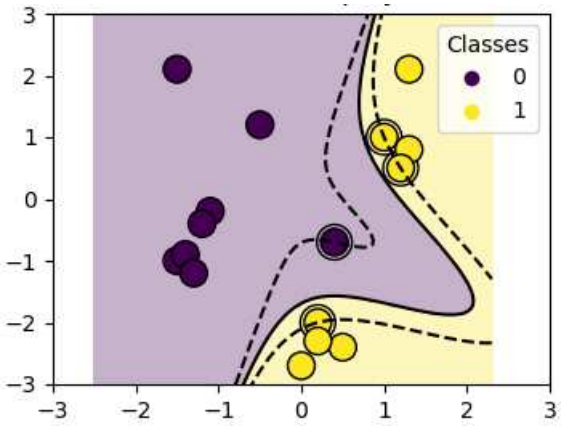
classificar dados em um conjunto de dados. Diferentes tipos de *kernel* podem ser utilizados em modelos SVM, resultando em tipos de funções diferentes utilizadas na modelagem dos hiperplanos, como o linear, exemplificado na Figura 2.3, o polinomial, apresentado na Figura 2.4, e a função de base radial (RBF, do inglês *Radial Basis Function*), mostrada na Figura 2.5.

Figura 2.3: Exemplo de hiperplano calculado em modelo SVM de *kernel* linear.



Fonte: (PEDREGOSA *et al.*, 2011).

Figura 2.4: Exemplo de hiperplano calculado em modelo SVM de *kernel* polinomial.

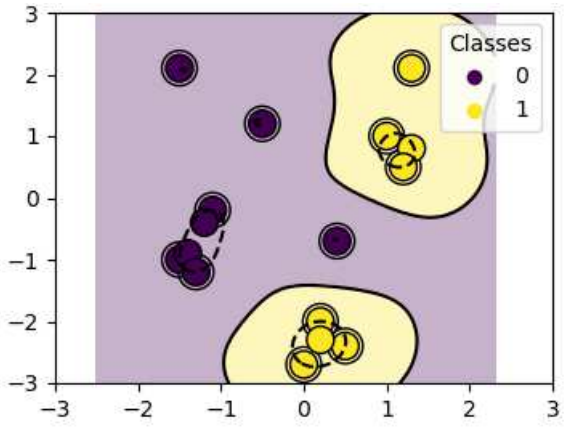


Fonte: (PEDREGOSA *et al.*, 2011).

Dentre os hiperparâmetros passíveis de aplicação ao modelo SVM, especialmente na implementação disponibilizada pela biblioteca sklearn[6], destacam-se: o C, o tipo de *kernel*, o grau ou *degree* (aplicável apenas para *kernel* polinomial, definindo o grau da função polinomial utilizada) e o *gamma*. O valor atribuído ao hiperparâmetro C reflete na regularização

[6]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Figura 2.5: Exemplo de hiperplano calculado em modelo SVM de *kernel* RFB.



Fonte: (PEDREGOSA *et al.*, 2011).

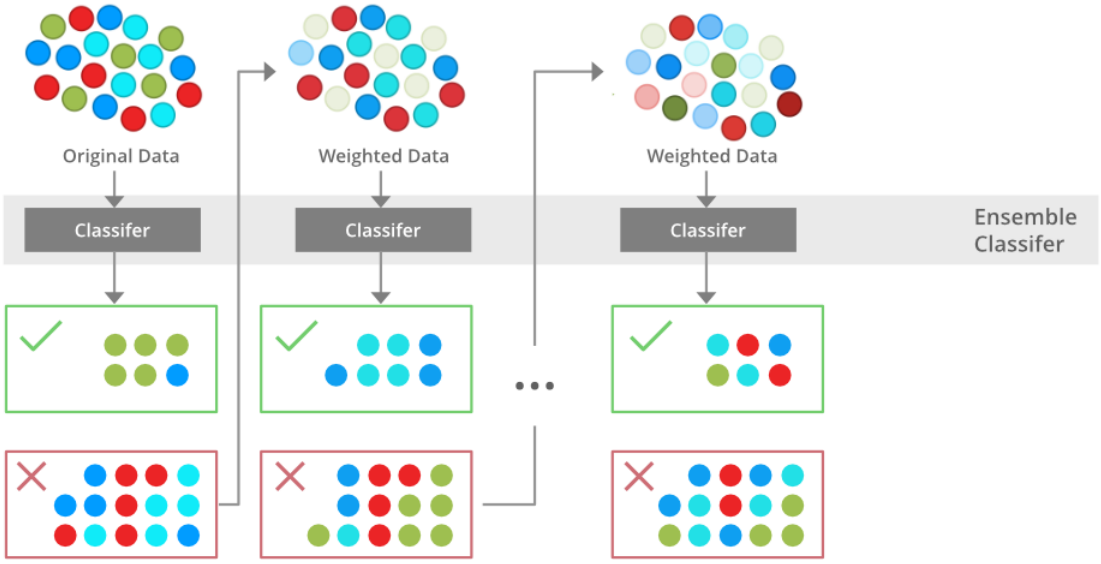
dos valores de erro nas regiões de fronteira, sendo inversamente proporcional ao seu impacto na regularização. Quanto ao *gamma*, este parâmetro define o fator de influência de cada instância única do conjunto de treinamento na fronteira de decisão, sendo um valor decimal. Na implementação do sklearn, o *gamma* também pode receber os valores "auto", onde o valor final de *gamma* é igual a $1/n_atributos$, ou "scale", onde *gamma* é definido como $1/(n_atributos * \sigma_x^2)$, em que $n_atributos$ representa a quantidade de atributos presentes em cada elemento da entrada X e σ_x^2 é a variância calculada para a entrada X completa.

2.4.2 XGBoost

O XGBoost (CHEN; GUESTRIN, 2016) é um algoritmo de AM do tipo *ensemble* amplamente aplicado em tarefas de classificação (QI, 2020; CHEN *et al.*, 2022a; HENDRAWAN; UTAMI; HARTANTO, 2022), combinando diferentes modelos para realizar a tarefa proposta. Baseado no conceito de aumento de gradiente, ou *gradient boosting*, o XGBoost utiliza árvores de decisão como os modelos base. Esse método busca mapear pequenas partes do problema completo em modelos menores que, quando combinados, culminam na resolução do problema completo. Conforme exemplificado na Figura 2.6, o *gradient boosting* utiliza classificadores, neste caso, árvores de decisão, para corrigir as classificações incorretas de seus predecessores.

Dentre os hiperparâmetros ajustáveis no XGBoost, temos: *booster*, *lambda*, *alpha*, *updater*, *top_k* e o *feature selector*. O *booster* define os tipos de modelos que serão utilizados

Figura 2.6: Fluxo de processamento de modelo *ensemble* utilizando técnica de *gradient boosting*.



Fonte: (XGBOOST..., 2023).

no XGBoost, dentre modelos baseados em árvores, com valores "gbtree" e "dart", ou modelos lineares, para o valor "gblinear". O impacto da regularização L2 aplicada ao modelo é definido pelo *lambda*, onde valores altos tornam o modelo mais conservador. De forma análoga ao *lambda*, o *alpha* impacta na regularização L1. O *updater*, define qual o algoritmo utilizado para ajustar os pesos do modelo no processo de treinamento. O hiperparâmetro *top_k* determina o número de atributos a serem selecionados pelo algoritmo, onde 0 significa utilizar todos os atributos disponíveis. Por fim, o *feature_selector* consiste no algoritmo de ordenação para seleção dos atributos utilizados pelo modelo. Outros hiperparâmetros específicos a cada tipo de *booster* estão listados na Tabela 2.1.

2.5 Modelos de Linguagem Grandes

Com o avançar dos estudos e agregação de novas tecnologias ao PLN, surgiram os Modelos de Linguagem Grandes (do inglês, *Large Language Models* - LLMs), sendo estes modelos de Aprendizagem Profunda (do inglês, *Deep Learning*). Até o surgimento dos *transformers*,

Tabela 2.1: Demais hiperparâmetros configuráveis no XGBoost.

Hiperparâmetro	***Booster***	**Descrição**
max_depth	"gbtree" ou "dart"	Profundidade máxima das árvores de decisão.
eta	"gbtree" ou "dart"	Tamano do *step* aplicado à redução de pesos nas atualizações para evitar *overfitting*.
gamma	"gbtree" ou "dart"	Redução mínima na perda para particionar um nó folha de uma árvore.
grow_policy	"gbtree" ou "dart"	Determina a forma como novos nós são adicionados às árvores.
sample_type	"dart"	Algoritmo de seleção de árvores.
normalizer_type	"dart"	Algoritmo de normalização aplicado às árvores.
rate_drop	"dart"	Taxa de *dropout* das árvores.
skip_drop	"dart"	Probabilidade de não realizar *dropout* durante a interação de *boosting*.

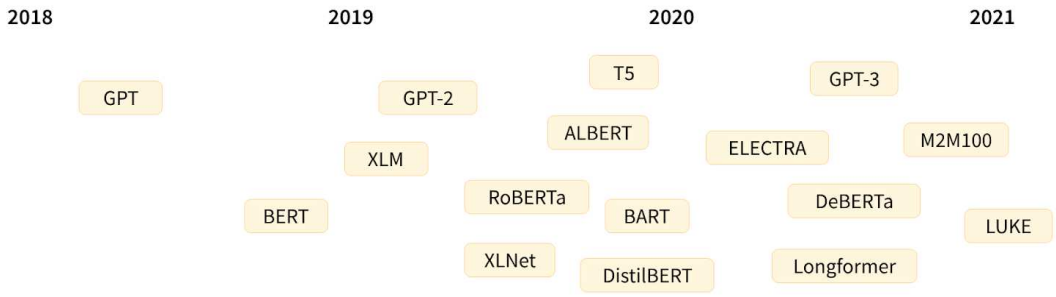
Fonte: De autoria própria, baseado na documentação do XGBoost (CHEN; GUESTRIN, 2016).

bem como seus modelos derivados, modelos utilizados em tarefas de PLN utilizavam-se apenas de conjuntos de dados anotados por meio de aprendizagem supervisionada, em um processo de treinamento custoso e altamente impactado pela qualidade e quantidade de dados (RADFORD *et al.*, 2018; LIU *et al.*, 2024). Diferentemente das técnicas tradicionais, os LLMs são pré-treinados em grandes volumes de dados textuais, utilizando-se de uma estratégia auto-supervisionada, o que resulta em modelos capazes de gerar, traduzir e reconhecer textos. Durante o processo de pré-treinamento, os modelos são capazes de reter os padrões de informações inerentes à linguagem, gravando uma grande quantidade de conhecimento linguístico em seus parâmetros (LIU *et al.*, 2024). Após a finalização do processo de pré-treinamento, todos os parâmetros mapeados pelos LLMs e o conhecimento adquirido podem ser utilizados para tarefas específicas, por meio de *fine-tuning*, utilizando-se de conjuntos de dados direcionados ou de abordagens de *few-shot learning* (LIU *et al.*, 2024).

Os LLMs surgiram a partir da criação do *transformer*, uma rede neural profunda considerada uma evolução frente aos modelos que utilizavam as complexas arquiteturas baseadas em redes neurais recorrentes e convolucionais (VASWANI *et al.*, 2017). A arquitetura *transformer* baseia-se no mecanismo de atenção, dispensando recorrências ou convoluções, seguindo uma estratégia *encoder-decoder*. Os *transformers* possibilitaram o surgimento de modelos

disruptivos, iniciando-se pela concepção do GPT, primeiro modelo de arquitetura baseada em *transformers*, adotando o formato de apenas *decoder* (RADFORD *et al.*, 2018; HANDI *et al.*, 2023; HOW..., [s.d.]). Logo em seguida surgiu o BERT, representante da arquitetura no formato de apenas *encoder* (DEVLIN *et al.*, 2019). Na Figura 2.7 está retratado o histórico de surgimento dos primeiros LLMs, partindo do ano de 2018 até 2021. Os modelos GPT e BERT atingiram resultados de estado da arte em diversas atividades de PLN, trilhando o caminho inicial ao que viria a ser uma revolução na área de PLN (HANDI *et al.*, 2023).

Figura 2.7: Referências dos anos de criação de modelos *transformer*.



Fonte: (HOW..., [s.d.])

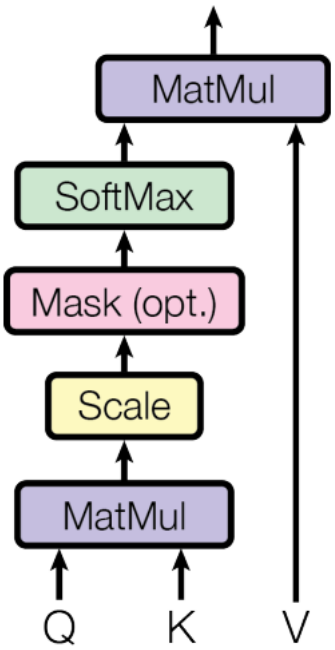
2.5.1 Atenção (*Attention* e *Multi-Head Attention*)

Os LLMs possuem como base o mecanismo de atenção, sendo o diferencial frente às soluções utilizadas anteriormente, como redes neurais recorrentes, por exemplo. Com a atenção, basicamente, busca-se responder a seguinte questão: "qual a parte da entrada que deve receber foco?". Ou seja, busca-se determinar o quão relevante é a palavra atual para as demais palavras da mesma sentença. Para cada palavra, é gerado um vetor de atenção que representa o relacionamento contextual entre as palavras da mesma sentença.

Na Figura 2.8 está disposto um diagrama representando a arquitetura de rede neural utilizada no cálculo da atenção, no formato de *single-headed attention*. A entrada é segmentada em três vetores: *query*, *key* e *value*. O cálculo do vetor resultante de atenção pode ser interpretado como a aplicação de uma busca (ou *query*) num conjunto de chaves e valores (*key* e *value*) (VASWANI *et al.*, 2017). O bloco *MatMul* inicial é responsável por realizar o produto escalar entre os vetores *query* e *key*, seguido pelo bloco *Scale* de normalização dos pesos pela

dimensão dos vetores *query* e *key*. A camada *Mask* oculta os valores de atenção referentes a palavras que ainda não ocorreram na sentença, sendo aplicado de forma opcional. É então aplicada a função *SoftMax* pelo bloco de nome análogo, partindo em seguida para o produto escalar do seu resultado pelo vetor *value*. Ao final, é obtida uma matriz de pesos que consiste no valor de atenção para cada palavra da sentença de entrada. A Equação 2.1 descreve o cálculo realizado a partir dos elementos dispostos na arquitetura da Figura 2.8, sendo Q, K e V as matrizes contendo os valores dos *query*, *key* e *value*, respectivamente, em que Q e K possuem dimensão d_k e V dimensão d_v.

Figura 2.8: Diagrama da arquitetura da rede neural utilizada no cálculo da atenção ou *single-headed attention* (*Scaled Dot-Product Attention*.)



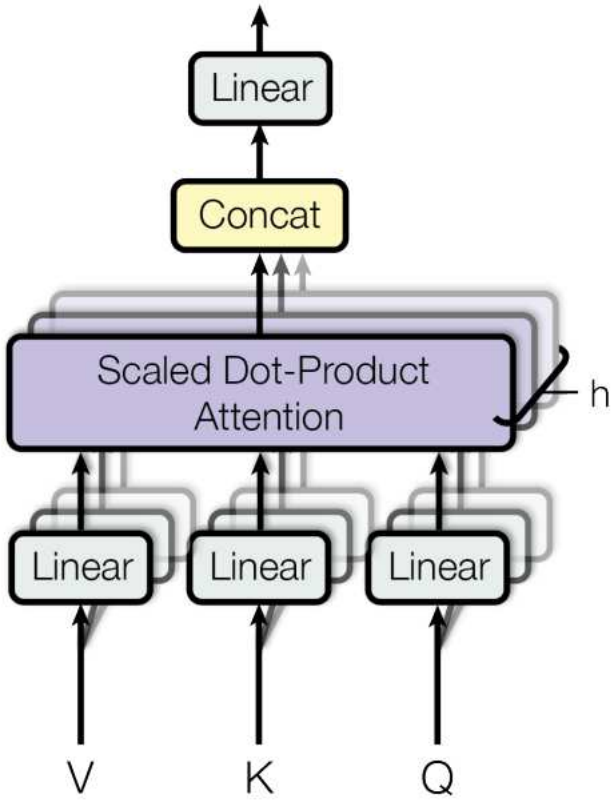
Fonte: (VASWANI *et al.*, 2017)

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

De forma a paralelizar o cálculo dos vetores de atenção, a entrada *query*, *key* e *value* pode ser segmentada em diferentes vetores de dimensões menores, realizando o cálculo da atenção. Ao final, os vetores são concatenados, conforme disposto no diagrama de *Multi-headed attention* na Figura 2.9, possuindo h vetores de atenção sendo calculados paralelamente e

concatenados ao final. A entrada é fragmentada utilizando-se uma matriz de pesos para cada "cabeça" de atenção, sendo esta a função da camada *Linear*. Ao concatenar o resultado de cada "cabeça" de atenção, o resultado também é aplicado a um vetor de pesos, conforme observado na Figura 2.9 e nas Equações 2.2 e 2.3, que descrevem o cálculo do *Multi-headed attention*. Sendo W_i^Q, W_i^K e W_i^V, na Equação 2.3, o vetor de pesos W para cada vetor de entrada para a "cabeça" de atenção, bem como W^O, a matriz de pesos da camada *Linear* final.

Figura 2.9: Diagrama da arquitetura da rede neural utilizada no cálculo da *Multi-headed attention*.



Fonte: (VASWANI *et al.*, 2017)

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2.2}$$

$$head_i(Q, K, V) = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.3}$$

2.5.2 *Word Embeddings*

Outro conceito importante tanto em problemas de PLN como para a compreensão do funcionamento do transformer é o de *Word Embeddings*, que são representações numéricas de palavras, normalmente em uma representação vetorial (MANDELBAUM; SHALEV, 2016). Os *Word Embeddings* são aplicados em diversas tarefas de PLN, de forma a mapear conceitos abstratos de palavras em espaços vetoriais, tornando palavras passíveis de serem processadas matematicamente (WANG; ZHOU; JIANG, 2019). Exemplos de ferramentas e métodos para a aplicação de *Word Embeddings* incluem: Word2Vec[7], GloVe[8] e ELMo[9].

2.5.3 *Transformer*

O mecanismo de atenção, sobretudo o *Multi-headed attention*, consiste na principal evolução trazida pelos *transformers*, entretanto a arquitetura geral é composta por outros elementos, conforme evidenciado no diagrama arquitetural da Figura 2.10. Dentre os elementos existentes, são exemplos a obtenção de *word embeddings* da entrada, a codificação posicional, as camadas de adição e normalização e, por fim, as camadas de *feedforward*. A arquitetura *transformer* também pode ser segmentada em duas principais partes, o *encoder* e o *decoder*. Modelos *transformers* podem utilizar tanto a arquitetura completa como apenas uma das partes, empilhadas em múltiplas instâncias, de acordo com o objetivo pretendido aos modelos em questão. Por exemplo, o BERT e diversos modelos derivados dele utilizam apenas a parte *encoder* da arquitetura, diferentemente do GPT e seus sucessores, que aplicam apenas a parte *decoder*.

2.5.3.1 Arquitetura *Encoder*

A parte do *encoder* da arquitetura *transformer* é responsável por extrair informações relevantes a partir da entrada alimentada (NAVEED *et al.*, 2023). Em um cenário *encoder-decoder*, o *embedding* obtido como saída do *encoder* é alimentado ao *decoder* para então gerar o texto correspondente, por exemplo, o texto traduzido, considerando um contexto de tradução de

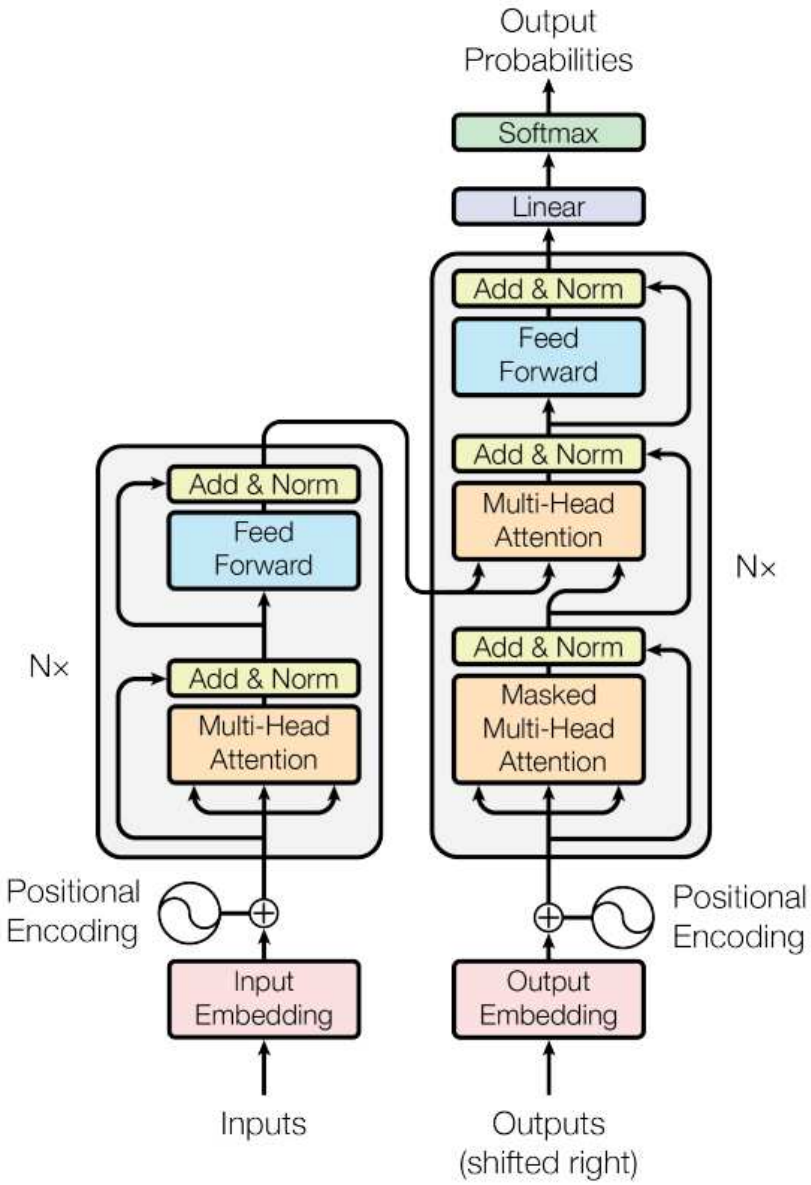
[7] https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057

[8] https://nlp.stanford.edu/projects/glove/

[9] https://allenai.org/allennlp/software/elmo

texto. Entretanto, modelos *encoder*, a exemplo do BERT, são pré-treinados utilizando-se de tarefas de pré-treinamento diferentes da de tradução, que foi originalmente utilizada como exemplo na concepção dos *transformers* (VASWANI *et al.*, 2017). Tarefas de pré-treinamento normalmente aplicadas aos modelos *encoder* direcionam os modelos a aprender representações contextuais a partir da entrada (DEVLIN *et al.*, 2019), tornando-os otimizados a tarefas de *Natural Language Understanding*, como: classificação de texto, análise de sentimentos e

Figura 2.10: Arquitetura de um *transformer*, exemplo possuindo formato *encoder-decoder*.



Fonte: (VASWANI *et al.*, 2017)

reconhecimento de entidades nomeadas (NAVEED *et al.*, 2023).

Para elucidar o funcionamento de um modelo de arquitetura *encoder*, tomemos como exemplo o modelo BERT. A tarefa de pré-treinamento aplicada a esse modelo, conhecida como Masked Language Model (MLM), consiste em mascarar aleatoriamente alguns dos *tokens* de entrada, substituindo-os pelo token [MASK]. Seu propósito é capacitar o modelo para prever as palavras mascaradas com base no contexto fornecido. Ao executar essa tarefa em um grande *corpus*, o modelo é capaz de modelar a linguagem e capturar informações em seus parâmetros, que são basicamente os pesos presentes nos elementos das redes neurais que compõem o modelo BERT. Vale destacar que o modelo BERT possui duas variantes: *base* e *large*, conforme detalhado em seus atributos listados na Tabela 2.2.

O tamanho oculto, ou *hidden size*, consiste nas camadas internas do modelo de funções matemáticas, atribuindo pesos às palavras de forma a atingir o resultado final. Em conjunto com a quantidade de camadas *transformer* e cabeças de atenção, o *hidden size* implica na quantidade de parâmetros existentes nos modelos. A entrada passada ao modelo é mapeada em *embeddings*, cada um com sua respectiva função, sendo eles: *tokens*, *segments* e *position*, conforme representado na Figura 2.11. Os *tokens* representam cada elemento da entrada em um espaço vetorial. Os *segments* determinam as sentenças as quais cada elemento da entrada pertence, dado que uma única entrada pode ser composta por diversas sentenças. Já os *positional embeddings*, são responsáveis por guardar a informação da posição dos elementos da entrada, localizando-os frente aos demais. A entrada é limitada pelo hiperparâmetro que determina o número máximo de *tokens* aceitos pelo modelo.

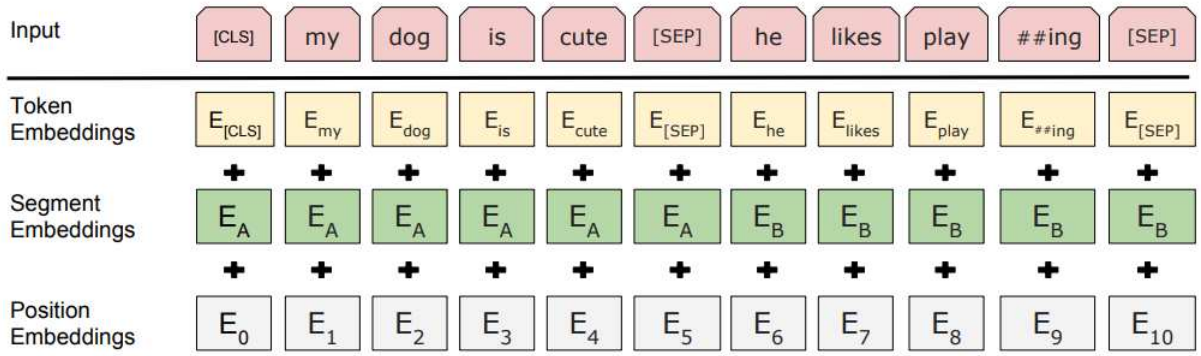
A utilização do modelo pré-treinado numa tarefa de classificação, por exemplo, consiste em acoplar uma camada de classificação ao final da arquitetura, mapeando assim a saída para o resultado esperado. A partir do processo de *fine-tuning*, que ajusta de maneira mínima o modelo de forma a adaptá-lo para a tarefa em questão, o modelo pode então mapear a sua

Tabela 2.2: Propriedades do modelo BERT em suas duas variantes (*base* e *large*).

	Camadas *Transformer*	Tamanho Oculto (*Hidden Size*)	Cabeças de Atenção	Parâmetros
BERT-base	12	768	12	110M
BERT-large	24	1024	16	340M

Fonte: De autoria própria, baseado na descrição do BERT (DEVLIN *et al.*, 2019).

Figura 2.11: Representação da entrada do BERT, detalhando todos os *embeddings* da entrada.

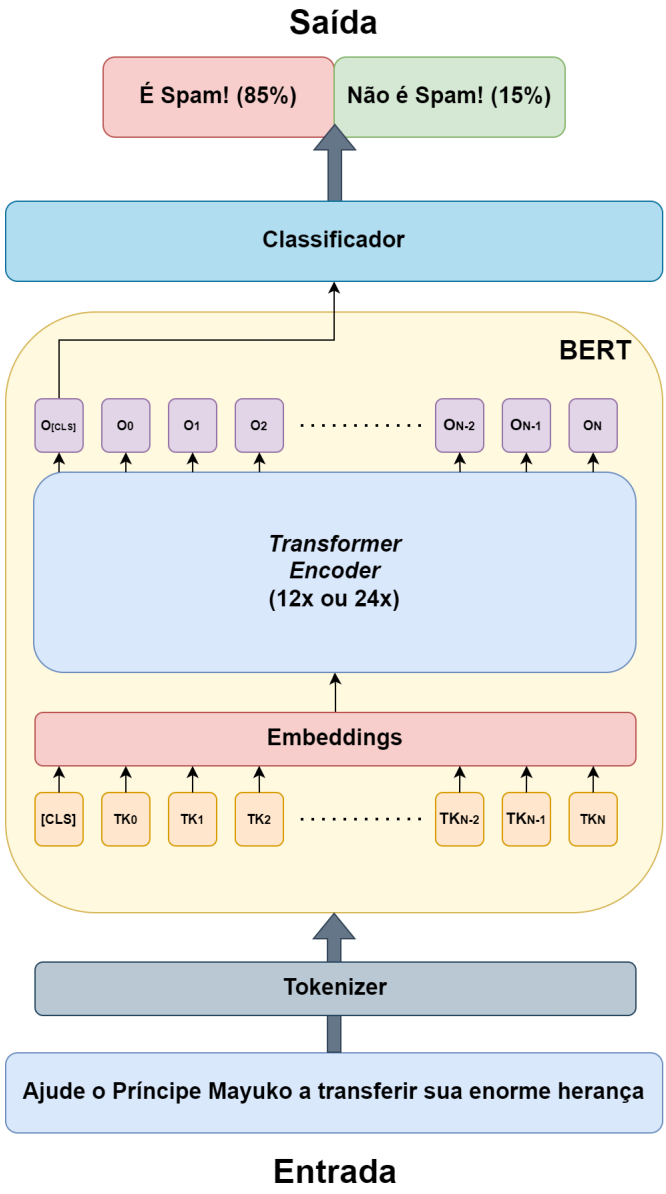


Fonte: (DEVLIN *et al.*, 2019).

saída para as classes alvo do problema. A Figura 2.12 contém o detalhamento do fluxo de utilização do BERT para um problema de classificação binária do título de um e-mail quanto a ser ou não spam. Inicialmente, o texto da entrada é convertido em *tokens*, dividindo as sentenças e palavras em unidades menores que serão utilizadas no processamento. Em seguida, são gerados os *embeddings*, bem como o vetor de atenção inicial a ser computado pelos *transformers*. Após percorrer todos os blocos *transformers*, a posição inicial da saída obtida é mapeada a um classificador ajustado ao problema em questão. Por fim, é obtida a saída em forma de probabilidade da classificação-alvo.

Dentre os diversos hiperparâmetros existentes passíveis de ajuste e otimização para modelos *transformer*, a taxa de aprendizagem (*learning rate*) e o tamanho dos lotes (*batch size*) foram hiperparâmetros relevantes considerados nesta pesquisa. O *learning rate* consiste no valor base para ajuste dos pesos do modelo durante o processo de treinamento, onde valores maiores implicam em ajustes mais drásticos e valores menores em ajustes mais suaves. Para o exemplo de *fine-tuning* do BERT, os autores utilizaram valores de *learning rate* da ordem de 10^{-5} (DEVLIN *et al.*, 2019). Quanto ao *batch size*, apresentando-se normalmente em valores da ordem de múltiplos de 2, consiste na quantidade de registros da entrada enviados ao modelo de forma paralela. Maiores valores de *batch size* implicam em menores tempos de treinamento, entretanto, num maior uso de memória por parte dos modelos.

Figura 2.12: Representação da entrada do BERT, exibindo todos os embeddings.



Fonte: De autoria própria baseado na descrição do BERT (ALAMMAR, 2018; DEVLIN *et al.*, 2019).

2.5.3.2 Arquitetura *Decoder*

Os modelos *decoder* são construídos utilizando-se da parte da arquitetura responsável pela geração de texto (NAVEED *et al.*, 2023). Sendo assim, os modelos que utilizam essa arquitetura caracterizam-se pela natureza generativa, a exemplo do GPT (RADFORD *et al.*, 2018). Sendo assim, tendem a ser direcionados para tarefas de *Natural Language Generation*, como criação de conteúdos e resposta a perguntas (NAVEED *et al.*, 2023). Além do tradicional *fine-*

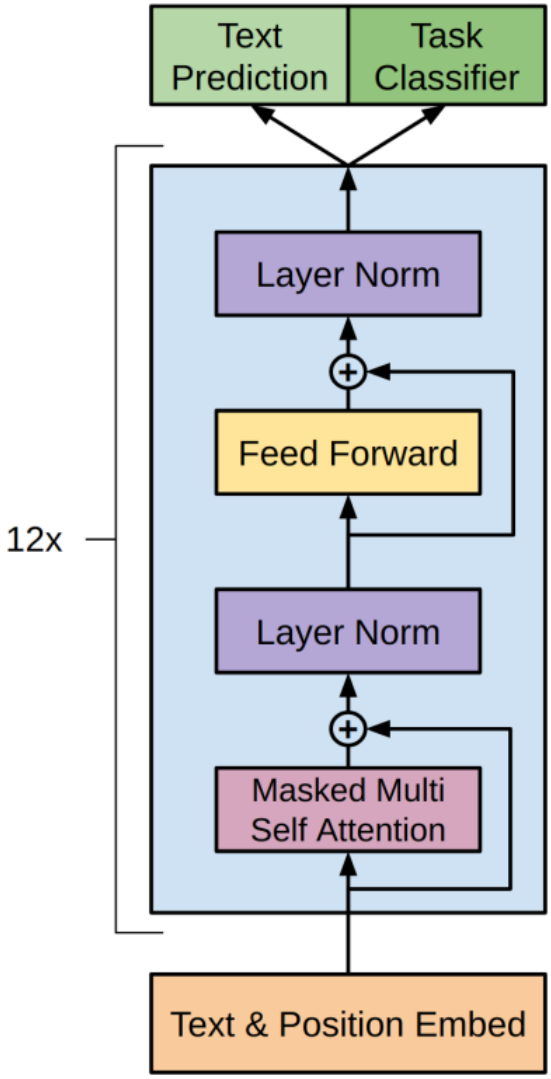
tuning, LLMs *decoder* podem ser utilizados em configurações chamadas *zero-shot*, *one-shot* ou *few-shot*, utilizando-se da janela de inferência dos modelos, aliadas aos seus parâmetros obtidos em pré-treinamento, para realizar tarefas com poucos ou nenhum exemplo (NAVEED *et al.*, 2023). Apesar de LLMs *decoder*, no geral, responderem bem a abordagens *few-shot*, esta forma de aprendizagem e inferência possui bastante influência do formato dos prompts utilizados no resultado (NAVEED *et al.*, 2023).

De forma a entender o funcionamento dos modelos de arquitetura *decoder*, é importante ter em mente o funcionamento do *Masked Multi-Head Attention*, primeira camada de atenção existente apenas no bloco *decoder*. Em resumo, o formato *masked* consiste em atribuir zero aos valores de atenção de cada *token* que ocorre após o *token* em questão. Por exemplo, serão computados os valores de relacionamento de uma palavra de uma determinada sentença para as outras, porém considerando apenas as palavras que ocorreram anteriormente a palavra alvo da análise. Desta forma, o modelo não considera *tokens* de palavras futuras, estando assim alinhado com a tarefa-alvo da arquitetura *decoder*, a de geração de texto. Um exemplo de tarefa de pré-treinamento utilizada em modelos *decoder*, tomando como base a tarefa utilizada no GPT (RADFORD *et al.*, 2018), é a de *language modeling*, que consiste em utilizar-se de um grande *corpus* de forma a fazer com que o modelo seja capaz de prever corretamente a próxima palavra de uma sentença. De forma semelhante ao observado no modelo BERT, a Figura 2.13 representa a arquitetura do modelo GPT, composta dos mesmos elementos do variante *encoder*, exceto pelo bloco de *Masked Multi-Head Attention* e o formato da saída, que consiste em texto.

A primeira versão do GPT possui aproximadamente 120 milhões de parâmetros (RADFORD *et al.*, 2018). Seus sucessores, GPT-2 e GPT-3, possuem, respectivamente, até 1,5 bilhões e 175 bilhoões de parâmetros (RADFORD *et al.*, 2019; BROWN *et al.*, 2020). Com o aumento do número de parâmetros dos LLMs generativos e o surgimento de novos modelos, o custo computacional necessário para realizar o *fine-tuning* para tarefas específicas também cresceu. Entretanto, percebeu-se uma maior quantidade de conhecimento retido, com modelos migrando de abordagens completamente baseadas em treinamento completo com ajustes de pesos para abordagens *few-shot*, valendo-se do conhecimento adiquirido pelos LLMs durante o pré-treinamento e *prompts* de contexto.

Outros LLMs *decoder* também foram lançados recentemente, como o Llama (TOUVRON

Figura 2.13: Representação da arquitetura do GPT, um LLM de arquitetura *decoder*.



Fonte: (RADFORD *et al.*, 2018).

et al., 2023a) e o Llama 2 (TOUVRON *et al.*, 2023b), construídos pela Meta[10], versão mais recente utilizando-se de um conjunto de treinamento contendo novos dados de fontes públicas. O Llama 2 consiste em um modelo aberto, com variantes de 7, 13, 34 e 70 bilhões de parâmetros, onde apresentou resultados competitivos frente a outros modelos em diversas tarefas de *benchmark*, dentre eles o próprio GPT-3. Sendo assim, o Llama, principalmente na sua versão 2, representa uma grande contribuição para o avanço de pesquisas na linha de PLN utilizando-se de LLMs *decoder*.

Além dos hiperparâmetros comuns a modelos de arquitetura *encoder*, como o *learning*

[10]https://llama.meta.com/

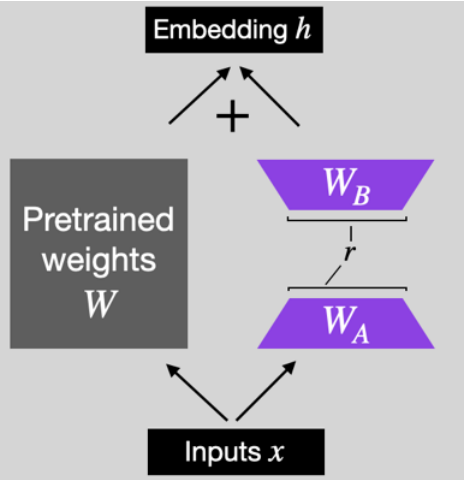
rate e o *batch size*, LLMs *decoder* recentes também incorporam hiperparâmetros direcionados especificamente ao processo de inferência, ou seja, à geração de texto (WANG; LIU; AWADALLAH, 2023; RENZE; GUVEN, 2024). Estes incluem: *temperature*, *top_p*, *top_k* e *max_new_tokens*. O hiperparâmetro *temperature*, com valor compreendido entre 0 e 1, controla a aleatoriedade do texto gerado, em que valores maiores implicam em textos mais aleatórios e diversos. O *top_p*, também com valor compreendido entre 0 e 1, controla a probabilidade de amostragem dos novos *tokens* gerados, em que valores baixos implicam na utilização de *tokens* com probabilidade mais alta, e valores altos resultam na exploração de uma possibilidade mais diversa de *tokens*. O *top_k*, um valor inteiro, direciona o modelo a considerar o token de saída como o de maior probabilidade dentre os primeiros *k tokens*. Tanto o *top_p* quanto o *top_k* são estratégias de amostragem de novos *tokens*, não podendo ser utilizados em conjunto. Por fim, o *max_new_tokens* determina a quantidade de *tokens* gerados como saída pelo LLM.

2.5.3.3 *Quantized Low Rank Adapters* (QLoRA)

Dada a crescente quantidade de parâmetros nos LLMs *decoder*, pesquisadores têm buscado estratégias para realizar o *fine-tuning* de forma mais eficiente e com menor consumo de memória. Uma dessas estratégias é o *Parameter-Efficient Fine-Tuning* (PEFT) (MANGRULKAR *et al.*, 2022), que consiste em métodos para adaptar os modelos de forma eficiente, ajustando apenas alguns parâmetros adicionais aos já existentes, o que reduz os custos e alcança resultados comparáveis aos do *fine-tuning* completo. O *Quantized Low Rank Adapters* (QLoRA) (DETTMERS *et al.*, 2023) consiste em uma das abordagens de PEFT existentes. Além de ajustar poucos parâmetros adicionais, o QLoRA aplica uma quantização aos modelos, reduzindo a precisão dos parâmetros para torná-los mais eficientes. Isso pode resultar em reduções para até inteiros de 4 bits. A Figura 2.14 apresenta a representação da estratégia aplicada pelo QLoRA aos pesos originais dos modelos.

A matriz de pesos W consiste nos pesos originais do modelo, que são congelados e não são ajustados. Duas novas matrizes menores de pesos W_A e W_B são adicionadas, onde apenas os pesos delas são ajustados durante o processo de *fine-tuning*. A saída h é determinada pela equação 2.4, que consiste na utilização dos pesos originais congelados W_0 em conjunto com os novos pesos adicionados W_A e W_B. O número de parâmetros treináveis é definido

Figura 2.14: Representação da estratégia de PEFT utilizada pelo QLoRA em LLMs.



Fonte: (RASCHKA, 2023)

por $|\Theta|$, calculado conforme a equação 2.5, onde $\hat{L}_{LoRA}$ consiste no número de matrizes nas quais a estratégia é aplicada, normalmente as matrizes *query* e *value*, d_{model} sendo a dimensão da matriz de pesos originais do modelo e r o *rank*, principal hiperparâmetro da estratégia e principal fator de influência no número de parâmetros treináveis.

$$h = W_0 x + \Delta W x = W_0 x + W_B W_A x \tag{2.4}$$

$$|\Theta| = 2 \times \hat{L}_{LoRA} \times d_{model} \times r \tag{2.5}$$

Além do *rank*, outros hiperparâmetros podem ser definidos na utilização do QLoRA, como *alpha*, *dropout* e *target_modules*. Quando os novos pesos ajustados são utilizados e adicionados aos pesos originais, são multiplicados por um fator de escala, determinado pelo valor de *alpha* dividido pelo *rank* r. Diminuiur os valores de *alpha* em comparação ao *rank* aumenta o impacto do *fine-tuning*. Aumentar o *alpha* em comparação ao *rank*, diminui o impacto do *fine-tuning* nos pesos originais. O *dropout* representa a probabilidade de que parâmetros treináveis serão atribuídos de zero para um dado lote de treinamento, sendo utilizado para evitar *overfitting*. O hiperparâmetro *target_modules* determina em quais pesos do modelo original a estratégia será aplicada, normalmente determinando-se os vetores *query* e *value*.

A quantização trata-se de um processo de discretização de uma entrada que possui mais informação para outra contendo menos informação (DETTMERS *et al.*, 2023). De forma a garantir que toda a escala da precisão alvo do processamento seja utilizada, os dados são ajustados utlizando-se normalização, por meio da utilização do máximo absoluto possível. Considerando uma redução de valores de ponto flutuante de 32 bits (FP32) para inteiros de 8 bits (Int8), o vetor irá possuir valores num intervalo de [-127, 127], sendo calculado conforme a equação 2.6 (DETTMERS *et al.*, 2023), em que c^{FP32} representa uma constante de conversão, de forma a simplificar a equação.

$$X^{Int8} = round\left(\frac{127}{absmax(X^{FP32})}X^{FP32}\right) = round(c^{FP32} \cdot X^{FP32}) \tag{2.6}$$

Dentre os hiperparâmetros passíveis de configuração para a quantização, ressaltam-se o método, que define a escala alvo da quantização, *compute_dtype*, que define a escala de conversão inicial da entrada passada ao modelo e, por fim, o *nested_quantization*, que consiste em habilitar a quantização em duas etapas, utilizando inicialmente uma escala intermediária antes da escala final (DETTMERS *et al.*, 2023).

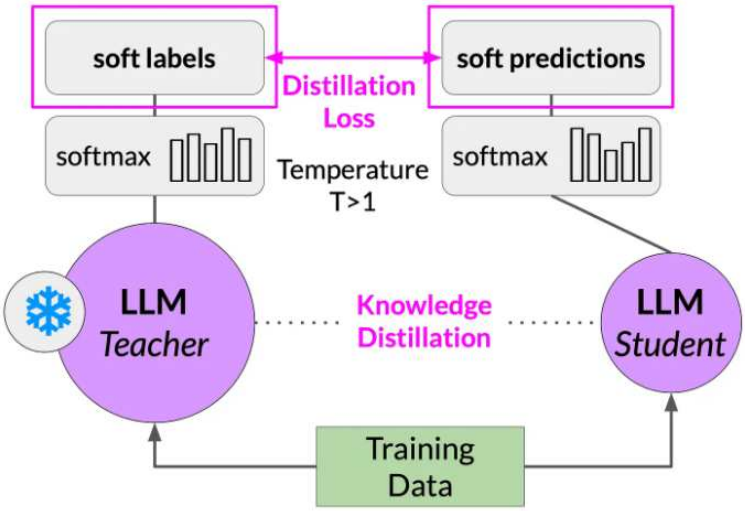
2.5.4 Modelos Destilados

Com o crescente aumento da demanda computacional para a utilização e treinamento de modelos, os pesquisadores buscaram alternativas para minimizar os impactos causados por esse inevitável caminho, como é o caso dos modelos destilados (GU *et al.*, 2023). De forma simplificada, a construção de modelos destilados, também conhecida como *Knowledge Distillation* (KD), consiste em transferir o conhecimento de um modelo maior, como os LLMs, para um modelo com menor quantidade de parâmetros. Essa transferência busca tanto maximizar os resultados obtidos com o modelo menor quanto alcançá-los de forma mais eficiente.

A estratégia utilizada para construir LLMs destilados, como é o caso do DistilBERT (SANH *et al.*, 2019), por exemplo, consiste em utilizar um modelo menor, chamado de estudante, em conjunto com um modelo maior, sendo este último o professor. O diagrama da Figura 2.15 representa a aplicação dessa estratégia. O processo deve iniciar-se com o mapeamento da saída de ambos os modelos para a mesma tarefa objetivo, de modo a alimentar ambos com a mesma entrada e comparar suas saídas para calcular o erro de destilação,

ou *distillation loss*. Em seguida, o modelo estudante é ajustado para minimizar esse erro, enquanto os pesos do modelo professor permanecem congelados. Ao final do processo, o modelo estudante resultante terá utilizado as saídas do modelo professor para ajustar-se e, consequentemente, aprender os pesos, buscando representar as mesmas informações com uma menor quantidade de parâmetros.

Figura 2.15: Representação da estratégia de *Knowledge Distillation* aplicada a LLMs.



Fonte: (SINGH, 2024)

2.6 Métricas de Avaliação

De forma a avaliar resultados obtidos mediante a aplicação de modelos de AM, a seleção de métricas de avaliação adequadas é essencial. Dentre as principais métricas utilizadas em problemas de PLN, bem como em AM em geral, destacam-se: acurácia, precisão, revocação e F1-Score (BLAGEC *et al.*, 2020). Considerando métricas focadas em problemas mais alinhados com PLN, são exemplos as métricas BLEU score, ROUGE e *Word Error Rate*. Em problemas de classificação, a construção de uma Matriz de Confusão auxilia na obtenção de percepções a partir da verificação de classificações corretas e incorretas. Considerando um cenário focado em classificação binária, destaca-se como uma métrica relevante a Área Sob a Curva de Precisão-Revocação (PR-AUC) (BOYD; ENG; PAGE, 2013).

Nesta seção, serão detalhadas as métricas relevantes ao problema de classificação de texto, em especial, em um cenário com apenas duas classes, ou seja, uma classificação biná-

ria, alvo da pesquisa descrita neste trabalho. As métricas selecionadas são: precisão, revocação, F1-Score, e PR-AUC. Também será discorrido acerca da matriz de confusão, estrutura de dados necessária para a obtenção das métricas selecionadas.

2.6.1 Matriz de Confusão

Auxiliando na representação sumarizada das predições de uma tarefa de classificação, a matriz de confusão consiste em um resumo em forma de matriz dos resultados obtidos em termos de quantidade de predições corretas e incorretas. As predições são fragmentadas em quatro diferentes tipos: Verdadeiros Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) e Verdadeiros Negativos (VN) (TIWARI, 2022). Matrizes de confusão podem ser aplicadas tanto em problemas de classificação binária como em múltiplas classes. Além de trazer um resumo quantitativo dos resultados obtidos, as células da matriz de confusão são utilizadas no cálculo de demais métricas, como precisão, revocação e F1-Score. Na Figura 2.16 é apresentada como se dá a disposição de uma matriz de confusão para cenários de classificação binária.

Figura 2.16: Disposição das células de uma matriz de confusão em um problema de classificação binária.

		Classe Obtida	
		P	**N**
Classe Verdadeira	**P**	VP	FN
	N	FP	VN

Fonte: De autoria própria.

Para o problema abordado neste trabalho, que foi modelado como uma classificação binária, a classe positiva refere-se aos atos tributários, enquanto a classe negativa diz respeito aos atos não tributários. Ao relacionar essas classes com os tipos de predições de uma matriz de confusão, temos que os VP representam os atos tributários corretamente identificados, enquanto os VN são os atos não tributários que também foram corretamente classificados. Por

outro lado, as classificações incorretas são representadas pelos FN, que são os atos tributários incorretamente classificados como não tributários, resultando em sua não identificação. Os FP, por sua vez, correspondem aos atos não tributários erroneamente apontados como tributários. No contexto específico do problema em análise, o impacto dos FP é considerado menor do que o dos FN. Isso ocorre porque é preferível ter alguns FP a deixar de capturar registros de interesse, como no caso dos FN.

2.6.2 Precisão

Considerando-se um cenário de classificação binária, onde as classes-alvo resumem-se a positiva ou negativa, é possível interpretar a precisão em termos de probabilidade. Ou seja, a precisão consiste na probabilidade de um registro classificado como positivo ser realmente positivo (GOUTTE; GAUSSIER, 2005). Altos valores de precisão indicam um cenário com uma baixa quantidade de FP, enquanto valores baixos de precisão representam um cenário com uma maior incidência de FP. Formalmente, a métrica consiste na quantidade de VP (registros classificados como positivos e que realmente são positivos frente às anotações) dividida pela quantidade de VP somada aos FP (registros classificados como positivos, mas que são realmente negativos) (GOUTTE; GAUSSIER, 2005). A Equação 2.7 descreve a fórmula para a precisão p.

$$p = \frac{VP}{VP + FP} \tag{2.7}$$

2.6.3 Revocação

Quanto a revocação (do inglês, *recall*), a mesma pode ser interpretada como a probabilidade de um registro positivo ser corretamente identificado pela modelagem. Altos valores de revocação indicam um cenário com uma baixa quantidade de FN, enquanto valores baixos de revocação representam um cenário com uma maior incidência de FN. Em termos formais, a revocação trata-se da quantidade de VP dividida pela quantidade de VP somada à quantidade de FN (registros positivos, mas que foram classificados como negativos) (GOUTTE; GAUSSIER, 2005). A Equação 2.8 descreve a equação para a revocação r.

$$r = \frac{VP}{VP + FN} \tag{2.8}$$

2.6.4 F1-Score

Para consolidar tanto a precisão quanto a revocação em uma única métrica, recorre-se ao F1-Score. Essa métrica consiste na média harmônica da precisão e da revocação, conforme a Equação 2.9. O F1-Score também pode ser descrito como Fβ-Score de valor $F\beta$, com β representando o fator de importância da revocação sobre a precisão (GOUTTE; GAUSSIER, 2005). Ou seja, como β = 1 para o F1-Score, ambas as métricas possuem o mesmo peso, chamada assim de média harmônica. o $F\beta$ pode ser calculado conforme descrito na Equação 2.10.

$$F1\text{-}Score = 2\frac{p * r}{p + r} \tag{2.9}$$

$$F\beta\text{-}Score = (1 + \beta^2)\frac{p * r}{(\beta^2 * p) + r} \tag{2.10}$$

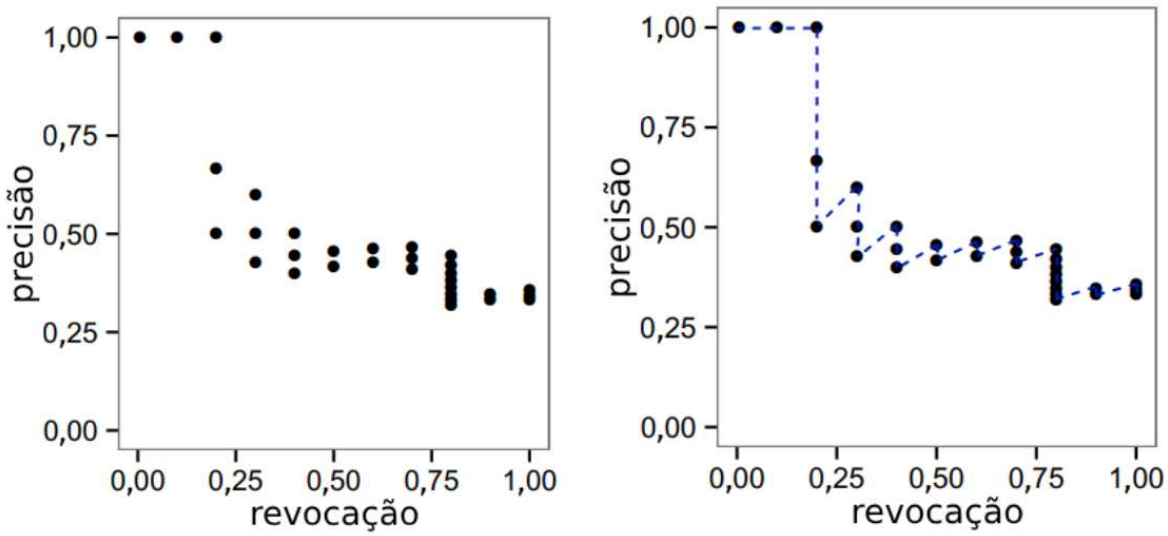
2.6.5 Área Sob a Curva de Precisão-Revocação

A curva de Precisão-Revocação (PR) é uma métrica de avaliação em contextos de classificação binária, tornando possível a visualização de desempenho em um intervalo de limiares de classificação (BOYD; ENG; PAGE, 2013). A curva de PR consiste numa representação bastante utilizada pela comunidade de AM, sobretudo em problemas envolvendo grandes desbalanceamentos nos dados (BOYD; ENG; PAGE, 2013; QI *et al.*, 2021). A utilização dessa representação gráfica da qualidade de classificadores binários não se limita a uma análise visual, também sendo calculada a Área Sob a Curva (do inglês, *Area Under the Curve*, ou AUC), onde a área resultante é chamada de PR-AUC. Assim, quanto maior for o valor da área abaixo da curva plotada, maior a qualidade do classificador analisado.

Para se construir uma curva de PR, faz-se necessário selecionar um conjunto de limiares de classificação para o problema de classificação binária em questão, atribuindo rótulos aos registros no conjunto utilizado para avaliar os resultados. Para cada limiar definido, os rótulos são obtidos e os valores de precisão e revocação calculados. Ao final, plota-se no gráfico

um ponto correspondente aos valores obtidos para cada um dos limiares utilizados. A Figura 2.17 contém um gráfico construído a partir de um conjunto de dados pequeno de exemplo, representando os pontos de PR, juntamente do gráfico contendo os pontos interligados, representando a curva de PR.

Figura 2.17: Gráfico de Exemplo de pontos de Precisão-Revocação para um conjunto de dados pequeno (à esquerda, apenas pontos obtidos a partir dos limiares de classificação; à direita, pontos interligados formando a Curva de PR).



Fonte: (BOYD; ENG; PAGE, 2013)

Em resumo, a métrica PR-AUC pode ser definida como a integral definida da precisão p em função da revocação r, conforme disposto na Equação 2.11 (KEILWAGEN; GROSSE; GRAU, 2014). Podem ser utilizados diversos métodos para o cálculo do valor final, como por exemplo, o método dos trapézios (BOYD; ENG; PAGE, 2013), que consiste em segmentar a área do gráfico em diversos trapézios (um para cada par de pontos contíguos), calcular isoladamente sua área e, em seguida, somar todas as áreas calculadas. O cálculo da PR-AUC utilizando-se do método dos trapézios está detalhado na Equação 2.12, em termos da precisão p, revocação r e n pontos existentes no gráfico da curva de PR.

$$PR\text{-}AUC = \int_0^1 p(r)dr \tag{2.11}$$

$$PR\text{-}AUC = \sum_{i=1}^{n-1} \frac{(r_{i+1} - r_i) * (p_{i+1} + p_i)}{2} \tag{2.12}$$

2.7 Considerações Finais

Foram apresentados neste capítulo os conceitos utilizados para o desenvolvimento do estudo, de forma a apoiar o entendimento dos próximos capítulos. Serão aplicadas técnicas de AM voltadas para PLN no domínio do direito tributário sob o contexto do DOU, temáticas discutidas neste capítulo. Também foram detalhadas as métricas essenciais para a avaliação dos resultados, levando-se em conta o contexto do problema aqui abordado.

No próximo capítulo são detalhados os trabalhos relacionados ao problema alvo desta pesquisa, de forma a situar o trabalho realizado frente a pesquisas existentes, buscando lacunas e contribuições passíveis de serem realizadas.

Capítulo 3

Trabalhos Relacionados

Neste capítulo serão discutidos os trabalhos relacionados ao tema desta pesquisa, de forma a situar o trabalho descrito neste documento frente ao estado da arte. Foi realizada uma busca por pesquisas com temáticas alinhadas com os objetivos e questões de pesquisa definidos. Sendo assim, foram levantados trabalhos nas linhas de LLMs, bem como seu pré-treinamento em domínios específicos, aplicações de PLN com relação ao domínio tributário, contextos de classificação altamente desbalanceados e discussões em torno de LLMs de arquitetura *encoder* e *decoder* com relação ao seu custo-benefício.

O presente capítulo é dividido da seguinte forma: na Seção 3.1 são descritos os avanços com relação aos LLMs pré-treinados na língua portuguesa; na Seção 3.2 são detalhados os avanços das pesquisas na linha da aplicação de PLN no domínio jurídico; na Seção 3.3 o foco foi direcionado para o PLN aplicado ao domínio tributário; na Seção 3.4 são contempladas as pesquisas referentes ao alto desbalanceamento de *corpora* no contexto jurídico; na Seção 3.5 são apontadas pesquisas acerca da comparação dentre LLMs *encoder* e *decoder*; finalmente, na Seção 3.6, são realizadas as considerações finais do capítulo acerca da análise das pesquisas e estado da arte em torno da temática tratada neste trabalho.

3.1 Modelos de Linguagem Grandes em Português

Com o surgimento do modelo BERT, um modelo disruptivo em sua época de lançamento e de arquitetura baseada em *transformers*, foram proporcionados resultados de estado da arte em diversas tarefas de *benchmark* para PLN (DEVLIN *et al.*, 2019). A aplicação crescente

de modelos *transformer* em problemas do mundo real envolvendo PLN tornou-se cada vez mais frequente, culminando no desenvolvimento de LLMs com as capacidades observadas nos dias atuais.

Estratégias de adaptações de modelos *transformers* para domínios específicos ou outros idiomas além do inglês, língua do *corpus* aplicado na tarefa de pré-treinamento original do BERT, foram ferramentas primordiais para o avanço contínuo de pesquisas de PLN, especialmente em português com o surgimento de um modelo baseado em *transformers* para o idioma. Observou-se que, apesar de haver uma versão multilíngue do BERT, tarefas em idiomas específicos diferentes do inglês poderiam beneficiar-se da aplicação de um pré-treinamento utilizando um extenso *corpus* no idioma alvo (RÖNNQVIST *et al.*, 2019; VIRTANEN *et al.*, 2019; NOZZA; BIANCHI; HOVY, 2020), dado que possíveis limitações na representação em línguas com um menor *corpus* no pré-treinamento podem apresentar representações inferiores quando comparado com a língua de maior representação no conjunto (PIRES *et al.*, 2022; NOZZA; BIANCHI; HOVY, 2020). A partir de então, pesquisadores trabalharam na criação do BERTimbau, modelo baseado no BERT e pré-treinado num extenso *corpus* em português (SOUZA; NOGUEIRA; LOTUFO, 2020). Seu surgimento impulsionou a área de pesquisa da aplicação de PLN em português, trazendo resultados de estado da arte em diversos domínios de aplicação (ARAÚJO *et al.*, 2023).

Recentemente, pesquisadores continuando na busca por aprimorar o arsenal de ferramentas passíveis da aplicação de PLN em português, culminaram no desenvolvimento de um novo modelo pré-treinado em português de arquitetura *transformer*, o ALBERTINA (RODRIGUES *et al.*, 2023). Baseado no DeBERTa (HE *et al.*, 2021), um LLM de arquitetura *encoder*, assim como o BERT, o novo modelo atingiu resultados resultados de estado da arte nos experimentos realizados pelos autores frente ao seu principal concorrente, o BERTimbau. Possuindo aproximadamente três vezes mais parâmetros do que o BERTimbau, o ALBERTINA consiste numa contribuição extremamente significativa para a inovação e aplicação de técnicas de PLN no idioma português.

Soluções de PLN baseadas em BERTimbau para o português continuam sendo empregadas num amplo leque de aplicações de modelagem do idioma com foco em tarefas de aprendizagem. O surgimento do ALBERTINA contribui para o avanço das pesquisas e experimentos envolvendo a língua portuguesa, traçando um novo caminho para a aplicação de

soluções cada vez mais avançadas nos diversos domínios de aplicação em meio ao idioma, além do já explorado inicialmente utilizando-se do BERTimbau.

3.2 Processamento de Linguagem Natural no Domínio Jurídico

Apesar da atenção crescente ao domínio jurídico em português (MARTINS; SILVA, 2021), usado no DOU, mais pesquisas neste idioma ainda se fazem necessárias para alcançar os avanços já obtidos no domínio jurídico em língua inglesa. No entanto, o domínio jurídico oferece um caminho natural para implementar técnicas avançadas de PLN, dada sua terminologia jurídica intricada, léxico especializado que difere da linguagem comum e o amplo volume de documentos textuais originados de diversos contextos legais.

O domínio jurídico abrange vários tipos de documentos textuais, incluindo contratos, legislação, processos judiciais e acordos de confidencialidade. Consequentemente, apesar de pertencerem ao mesmo domínio, esta especificidade pode ser fator de influência na aplicação de técnicas de PLN em cenários reais. Assim, não é possível assegurar que um modelo pré-treinado em um *corpus* centrado principalmente em textos de licitações, por exemplo, teria o mesmo desempenho quando aplicado a um *corpus* de processos judiciais.

Nos últimos anos, alguns pesquisadores deram passos significativos na classificação de textos jurídicos (CLAVIÉ; ALPHONSUS, 2021; NGHIEM *et al.*, 2022; NGUYEN *et al.*, 2022; MAMOOLER *et al.*, 2022; CHEN *et al.*, 2022b). Outro domínio de pesquisa em evidência envolve a classificação e o processamento de documentos oficiais (SENGUPTA; DAVE, 2021; CAVALIERI *et al.*, 2022). Esses documentos, conhecidos por sua natureza textualmente densa, oferecem um cenário desafiador para a aplicação de técnicas de extração, processamento e classificação.

Entre os esforços de pesquisas recentes no domínio jurídico brasileiro, destacam-se algumas contribuições como VICTOR (ARAUJO *et al.*, 2020) e BERTIKal (POLO *et al.*, 2021). O VICTOR é um conjunto de dados focado na classificação de documentos jurídicos, composto principalmente de documentos do Supremo Tribunal Federal. BERTikal é um modelo de linguagem, construído sobre o BERTimbau, treinado em um *corpus* de textos jurídicos do Tribunal de Justiça do Estado de São Paulo. Essas contribuições desempenharam um pa-

pel fundamental em abordar a lacuna de pesquisa na análise de textos jurídicos brasileiros, fornecendo aos pesquisadores de PLN ferramentas para o desenvolvimento de modelos e incentivando o surgimento de pesquisas relevantes adicionais.

Exemplificando contribuições derivadas desses estudos, pode-se destacar: a aplicação de modelagem de tópicos na classificação de temática de processos, com casos oriundos da Suprema Corte (ARAUJO; CAMPOS, 2020) e da Corte de Justiça do Ceará (AGUIAR *et al.*, 2022); a extração de jurisprudência baseada em similaridade textual (GOMES; LADEIRA, 2020); e a recuperação de informações por meio de perguntas e respostas em documentos jurídicos (GOMES; LADEIRA, 2020).

A abordagem mais relevante devido a proximidade com os objetivos de pesquisa deste trabalho de mestrado envolve o monitoramento de alterações legislativas ambientais publicadas no DOU (CAÇÃO *et al.*, 2022). Nessa pesquisa, os autores concentraram-se na construção de um *corpus* dedicado ao domínio do DOU. Esse *corpus* foi utilizado pelos autores como base para o pré-treinamento do modelo BERDOU, construído sobre o BERTimbau. O BERDOU tem o potencial de aprimorar significativamente a aplicação de técnicas de PLN no contexto do DOU. Além disso, os autores criaram um conjunto de dados anotados adaptado para a tarefa de classificação de texto de leis ambientais, tendo sido direcionado a uma abordagem multiclasse.

3.3 Processamento de Linguagem Natural Aplicado ao Domínio Tributário

A aplicação de PLN ao contexto do domínio tributário ainda requer avanços na literatura, dado que é um domínio ainda pouco explorado de forma isolada, sendo muitas vezes atrelado como subdomínio em contextos jurídicos (ASH; GUILLOT; HAN, 2021; GU *et al.*, 2022). Não obstante, há algumas pesquisas recentes.

Uma das pesquisas levantadas com foco no domínio tributário endereçou a extração de leis tributárias a partir de textos legislativos dos Estados Unidos (ASH; GUILLOT; HAN, 2021). A pesquisa em questão utilizou um *corpus* pré-existente, compreendido entre o período de 1910 e 2010, e reforçou a natureza altamente desbalanceada do contexto, dado que apenas 6,87% dos registros presentes são tributários. Além da classificação binária, os autores ende-

reçaram a classificação de texto tributário dentre seus subdomínios, como: tributos corporativos, tributos sob energia, tributos sob heranças, dentre outros. Para a classificação binária, os autores utilizaram Regressão Logística e *Random Forests*, ambas abordagens tradicionais de aprendizagem de máquina. Apesar das métricas de AUC ROC para ambos os modelos serem altas, o F1-Score, juntamente da revocação, apresentaram valores em torno de 0,5 e 0,4, respectivamente. Salienta-se que, um baixo valor de revocação, implica em registros tributários que não foram identificados pelo modelo de aprendizagem, um ponto crítico na análise automatizada desses registros.

Sendo assim, resultados podem ter sofrido influência do alto desbalanceamento dos dados, gerando viés a partir do conjunto de treinamento empregado. Tal característica do conjunto não foi explorada pelos autores, além de uma breve descrição do conjunto de dados já existente e utilizado. Além disso, a natureza temporal dos dados foi ignorada, dada a aplicação de validação cruzada do tipo *five-fold*, havendo uma mistura dentre os registros de diferentes anos. Ou seja, foram utilizados registros de 2010 para treinamento do modelo, e de 1910 no conjunto de testes, por exemplo, o que impacta na confiabilidade dos resultados na aplicação dos modelos obtidos num cenário real em dados ainda não vistos.

Outro trabalho endereçou o domínio tributário no idioma coreano (GU *et al.*, 2022), inclusive com modelos baseados em arquitetura *transformer*. A aplicação do estudo consiste na classificação de perguntas enviadas a uma plataforma governamental para retirada de dúvidas relacionadas à legislação tributária. O cidadão que envia perguntas deve selecionar a temática de interesse, de forma que seja enviada a um analista especialista em direito tributário. Porém, segundo os autores, a maioria dos indivíduos seleciona as categorias de forma incorreta, numa proporção não informada. Sendo assim, os autores buscaram utilizar-se da classificação do texto da pergunta de forma a validar a temática selecionada para cada pergunta e direcioná-la ao analista tributário correto, evitando assim um redirecionamento de perguntas de um analista para outro.

Os autores realizaram um estudo comparando estratégias tradicionais de AM, como LSTM e BiGRU-CNN, além de LLMs já existentes, inclusive em coreano. Além disso, foi realizado o pré-treinamento de um novo modelo desde o início, de forma a comparar com os resultados dos demais modelos. Os autores obtiveram melhores resultados com o modelo por eles construído, aproximadamente 1% melhor em termos de F1-Score quando comparado

com o segundo melhor modelo, o KcBERT, modelo BERT pré-treinado na língua coreana. Apesar disso, o ganho não foi tão expressivo diante do elevado custo de pré-treinamento de um novo modelo, mesmo sendo um modelo destilado (SANH *et al.*, 2019; GU *et al.*, 2023), baseado no DistilRoBERTa[1]. O custo total de pré-treinamento do novo modelo em termos de tempo de processamento não foi discutido pelos autores, nem levado em consideração na comparação com os demais modelos.

Além disso, o problema-alvo apontado pelos autores, além de poder ser atendido com o *fine-tuning* de modelos já existentes, não remove a interação com o analista tributário alvo da pergunta, que precisa ser respondida. Adicionalmente, não é descrito pelos autores a proporção de usuários que realizaram a classificação errada da temática de suas perguntas. Esses fatores representam pontos que carecem melhor esclarecimento para auxiliar na justificativa e na relevância do problema atendido.

Por fim, não foram encontrados estudos relevantes que abordam especificamente o domínio tributário em português brasileiro, especialmente com foco no contexto dos Diários Oficiais e na identificação automática de publicações neste domínio.

3.4 Contextos Desbalanceados no Domínio Jurídico

Trabalhos de pesquisa recentes têm explorado a natureza desbalanceada dos textos jurídicos e as técnicas associadas de PLN. Isso ressalta a relevância e o desafio de lidar com o problema do desbalanceamento em tarefas de classificação de texto. Este problema impacta diretamente o contexto de classificação de publicações relacionadas ao domínio tributário no DOU, objeto de estudo nesta pesquisa.

Dentre os estudos analisados, uma pesquisa abordou estratégias específicas de amostragem de dados para minimizar os efeitos do desbalanceamento, inerente aos dados jurídicos, conforme indicado pelos autores (FREIRE *et al.*, 2023). Os experimentos foram conduzidos com documentos em português provenientes de processos do Tribunal de Justiça de São Paulo (TJSP), concentrando-se na classificação de processos quanto à temática de direito do consumidor. Para a modelagem, empregou-se o SVM com gradiente descendente estocástico, associado à vetorização usando TF-IDF e diversas estratégias de amostragem e

[1] https://huggingface.co/distilroberta-base

balanceamento de dados. Os resultados apresentaram uma média de F1-Score de aproximadamente 0,995, com um desvio padrão de aproximadamente 0,005, resultados extremamente próximos uns dos outros, podendo haver a influência da amostra aleatória em alguns métodos de amostragem nos resultados.

Contudo, os resultados não foram detalhados para cada classe separadamente no conjunto de testes, o que limita a compreensão do cenário. Apesar da contribuição da pesquisa na experimentação de diferentes abordagens de balanceamento e amostragem, os resultados deixam espaço para uma análise mais aprofundada desses aspectos no domínio jurídico.

Outro estudo, dessa vez focado em aplicar PLN na modelagem de casos da Suprema Corte dos Estados Unidos, também reforçou e buscou endereçar a natureza desbalanceada dos textos jurídicos (LOCKARD; SLATER; SUCRESE, 2023). A modelagem proposta pelos autores consistiu em utilizar um modelo de classificação, mais especificamente uma rede neural LSTM associada ao ELMo para vetorização do texto, de forma a determinar o resultado dos processos, ou seja, a parte vencedora: peticionário ou respondente. Os resultados obtidos pelos autores apresentaram um F1-Score de 0,32, tendo sido apresentado apenas a métrica englobando ambas as classes-alvo da tarefa, associado à um PR-AUC de 0,68, indicando que o modelo até certo ponto possui a capacidade de distinguir dentre as classes por estar acima de 0,5. Ainda assim, os resultados deixam margem para melhorias e análises futuras, como na obtenção de limiar de classificação que otimize o F1-Score, também na aplicação de diferentes estratégias de balanceamento dos dados de treinamento, nas estratégias de vetorização e na experimentação com outros modelos de classificação, principalmente LLMs.

Outros pesquisadores também abordaram o desbalanceamento de dados no domínio jurídico, empregando tanto técnicas tradicionais como modelos *transformers* em tarefas de classificação binária e multiclasse (IMRAN *et al.*, 2023). O conjunto de dados experimental continha casos do Tribunal Europeu de Direitos Humanos (TEDH)[2], onde para classificação binária as classes se dividem em “há violação” e “não há violação”, já para o problema multiclasse, cada classe corresponde ao tipo de violação aos direitos humanos, bem como a classe negativa “não há violação”. Para lidar com o desbalanceamento, especialmente no problema multiclasse, os pesquisadores obtiveram mais dados da classe "há violação" diretamente do site do TEDH. Os resultados indicaram que o modelo RoBERTa obteve o melhor F1-Score

[2]https://hudoc.echr.coe.int/

para a classificação binária (aproximadamente 0,87), enquanto o BigBird destacou-se para a classificação multiclasse (F1-Score de 0,78).

De forma a tratar o desbalanceamento presente no conjunto de dados, presente tanto no problema de classificação binária como no multiclasse, porém neste último ainda mais proeminente, os autores utilizaram-se da estratégia de obtenção de mais dados da classe "há violação" diretamente a partir do website do TEDH, com o objetivo de obter mais exemplos para cada um dos tipos de violação presentes na abordagem multiclasse. Entretanto, ao executar esta abordagem, os autores tornaram o cenário de classificação binária ainda mais desbalanceado. Nos experimentos realizados não houve qualquer comparação dos resultados antes e depois da inclusão das novas instâncias obtidas, o que limita a interpretação dos benefícios da expansão do conjunto de dados. Essas lacunas revelam a necessidade de investigações mais abrangentes.

Ao analisar os estudos, destaca-se a recorrência do desbalanceamento no domínio jurídico. Lidar com o desbalanceamento é desafiador, levando os pesquisadores a buscar soluções específicas para cada cenário de aplicação de PLN no domínio jurídico. No entanto, observa-se a existência de lacunas nas pesquisas analisadas, principalmente considerando o potencial impacto de desbalanceamentos negligenciados em problemas de PLN, como o viés. Aplicações no domínio jurídico frequentemente envolvem dados sensíveis e tarefas aplicadas em processos de decisão de alto impacto social, tornando a confiabilidade nos modelos um requisito crítico. Portanto, contribuições que se concentram no tratamento do desbalanceamento de dados, especialmente em proporções elevadas, são altamente relevantes para o domínio jurídico e para a área de PLN como um todo.

3.5 Modelos de Linguagem Grandes: *encoder* versus *decoder*

Pesquisas recentes têm contribuído para os avanços na aplicação de LLMs em tarefas de PLN, especialmente aqueles com arquitetura *decoder*, que obtiveram destaque ao longo do tempo (NAVEED *et al.*, 2023). Esses modelos, dotados de capacidade generativa, têm alcançado resultados de estado da arte em tarefas de PLN, apresentando quantidades de parâmetros na ordem das centenas de bilhões. A premissa é que esses modelos possuem uma

capacidade superior de representação e informação agregada, demandando nenhum ou mínimos ajustes para atividades específicas, aproveitando abordagens de aprendizagem *few-shot*. Contudo, o aumento constante da complexidade e quantidade de parâmetros desses modelos implica em maiores demandas computacionais para experimentação e, consequentemente, para a implementação em cenários práticos (VRIES, 2023).

Apesar da alta capacidade de representação e adaptação, os custos crescentes suscitam questionamentos sobre o custo-benefício da utilização de LLMs com arquitetura *decoder* em tarefas de NLP já abordadas por modelos tradicionais ou por LLMs com arquitetura *encoder*, como o BERT. Estudos recentes compararam essas abordagens, avaliando e respondendo a essas questões.

Em um desses estudos (ZHONG *et al.*, 2023), foi conduzida uma comparação entre o ChatGPT, em sua versão baseada no modelo GPT 3.5, e LLMs com arquitetura *encoder*, como o BERT e o RoBERTa, em suas versões *base* e *large*. Os experimentos abordaram tarefas de Compreensão de Linguagem Natural (do inglês, *Natural Language Understanding*, ou NLU), incluindo análise de sentimento, aceitabilidade linguística, detecção de paráfrase, similaridade de texto, implicação textual e implicação de perguntas e respostas. Para avaliar a capacidade dos modelos considerando seu pré-treinamento, foram realizados experimentos em tarefas de inferência. Os resultados indicaram que o ChatGPT foi superior nas tarefas que envolviam inferência, mas os LLMs com arquitetura *encoder* apresentaram desempenho comparável ou superior nas demais tarefas de NLU. Segundo os autores, o desempenho do ChatGPT em um cenário de *zero-shot* foi comparável ao do modelo BERT-base após o fine-tuning para as tarefas específicas. Apesar de destacar a notável capacidade de representação dos LLMs com arquitetura *decoder* para resolver problemas gerais, a pesquisa salienta que soluções específicas podem ser mais eficazes, tanto em termos de resultados quanto de custos, utilizando LLMs com arquitetura *encoder*.

Outra pesquisa relevante abordando esse tema (YU *et al.*, 2023), utilizando os modelos LLama 2 e GPT nas versões 3.5 e 4 como LLMs com arquitetura *decoder*, e o RoBERTa como representante da variante *encoder*, também comparou custo e desempenho, concentrando-se em atividades de classificação de texto e Reconhecimento de Entidades Nomeadas (do inglês, *Named Entity Recognition*, ou NER). Os resultados indicaram que modelos menores, aplicados em um contexto de classificação supervisionada, atingem de-

sempenho semelhante ou superior em comparação com os LLMs generativos. Além disso, para os experimentos realizados, o modelo RoBERTa exigiu uma fração do tempo necessário para treinamento e inferência em comparação com o LLama 2, que tinha 70 bilhões de parâmetros, sendo 5% e 1%, respectivamente.

Devido aos custos elevados da execução de experimentos utilizando os LLMs generativos, foi necessário fazer concessões e limitações em alguns pontos. O *fine-tuning* do modelo LLama 2, por exemplo, foi realizado com quantização e QLoRA (DETTMERS *et al.*, 2023), e o GPT-4 não foi aplicado em todos os conjuntos de dados. Os autores destacam que LLMs menores, como o RoBERTa, são alternativas viáveis aos enormes LLMs generativos, oferecendo resultados superiores com menor custo, melhor desempenho e transparência, especialmente em tarefas específicas. No entanto, aplicações de LLMs em tarefas que requerem generalização continuam apresentando melhores resultados quando se utiliza arquitetura *decoder*.

Por fim, a partir das pesquisas levantadas, observa-se que os LLMs de arquitetura encoder permanecem como alternativas viáveis e competitivas em aplicações de PLN, demonstrando desempenho semelhante ou superior em tarefas de classificação de texto e NLU. Além disso, modelos maiores implicam em um consumo significativamente maior de recursos, aumentando os custos e apresentando desafios, tanto para treinamento de modelos quanto na aplicação em cenários práticos.

3.6 Considerações Finais

A análise dos trabalhos relacionados proporcionou uma visão abrangente do estado da arte no campo da aplicação de PLN e LLMs nos contextos jurídico e tributário. No decorrer deste capítulo, foram discutidos avanços significativos em diversas linhas de pesquisa, cada uma abordando aspectos específicos que contribuem para o entendimento aprofundado do tema em questão.

O domínio jurídico, dada sua complexidade terminológica e vasto volume de documentos, apresenta-se como um terreno fértil para a implementação de técnicas avançadas de PLN. As investigações existentes na linha do domínio tributário, embora valorosas, mostraram-se limitadas em abrangência e enfrentam desafios, como desbalanceamento de dados e custos associados ao pré-treinamento de modelos específicos. Além disso, o domínio tributário

carece de abordagens envolvendo PLN, principalmente em português. Ao reunir essas análises, é possível perceber que a pesquisa no campo de PLN aplicada aos contextos jurídico e tributário está em constante evolução, mas desafios significativos ainda persistem.

Sendo assim, a partir da análise realizada, destacam-se as oportunidades para a aplicação de técnicas de AM de estado da arte, especialmente em PLN, com foco em classificação binária de texto tributário no contexto altamente desbalanceado do DOU, uma abordagem inovadora no domínio frente aos trabalhos analisados. A Tabela 3.1 contém a sumarização dos principais trabalhos e suas contribuições para a análise realizada.

No próximo capítulo, é detalhada a abordagem metodológica para a condução da pesquisa e endereçamento da problemática levantada, sendo discutidos os procedimentos, instrumentos e estratégias utilizados na coleta, modelagem e análise dos dados e resultados.

Tabela 3.1: Sumarização dos principais trabalhos relacionados levantados na análise realizada.

Referência	**Domínio**	**Técnicas Utilizadas**	**Resumo da Contribuição**
(SOUZA; NOGUEIRA; LOTUFO, 2020)	LLM em Português	BERT; BERTimbau	BERTimbau, LLM *encoder* em português baseado no BERT.
(RODRIGUES *et al.*, 2023)	LLM em Português	DeBERTa; ALBERTINA	ALBERTINA, LLM *encoder* em português, baseado no DeBERTa.
(POLO *et al.*, 2021)	PLN no Dom. Jurídico	BERTimbau; BERTIKAL	BERTIKAL, LLM de dom. jurídico, baseado no BERTimbau.
(CAÇÃO *et al.*, 2022)	PLN no Dom. Jurídico	BERTimbau; BERDOU	Classificação multiclasse de atos ambientais do DOU; BERDOU, LLM com foco no DOU.
(ASH; GUILLOT; HAN, 2021)	PLN no Dom. Tributário	Regressão Logística; *Random Forests*	Extração de leis tributárias a partir de textos legislativos dos EUA.
(GU *et al.*, 2022)	PLN no Dom. Tributário	LSTM; BiGRU-CNN; KcBERT; DistilRoBERTa; KTL-BERT	LLMs no domínio tributário; classsificação multiclasse; pré-treinamento de LLM no domínio tributário.
Pesquisa conduzida neste trabalho.	**PLN no Dom. Tributário**	SVM; XGBoost; PA; BERTimbau; Legal-BERT; BERTIKAL; BERDOU; XLM-RoBERTa; ALBERTINA; Llama2	**Classificação de atos tributários do DOU; comparação de proporções de balanceamento; LLMs *encoder* x *decoder*.**
(FREIRE *et al.*, 2023)	Desba. no Dom. Jurídico	SVM + TF-IDF	Classificação multiclasse em direito do consumidor no TJSP; estratégias para desbalanceamento.
(LOCKARD; SLATER; SUCRESE, 2023)	Desba. no Dom. Jurídico	LSTM + ELMo	Modelagem de casos da Suprema Corte dos EUA, determinando vencedor.
(IMRAN *et al.*, 2023)	Desba. no Dom. Jurídico	SVM; DT; NB; AdaBoost; BERT; Legal-BERT; RoBERTa; BigBird; ELECTRA; XLNet	Técnicas tradicionais e LLMs; classificação binária e multiclasse em casos do TEDH.
(NAVEED *et al.*, 2023)	LLMs *encoder x decoder*	Revisão sistemática de LLMs	Histórico de surgimento de LLMs, contribuições e desafios.
(ZHONG *et al.*, 2023)	LLMs *encoder x decoder*	ChatGPT (3.5); BERT; RoBERTA.	LLMs *encoder* x *decoder* em tarefas de NLU e de inferência.
(YU *et al.*, 2023)	LLMs *encoder x decoder*	GPT-3.5; GPT-4; Llama2; RoBERTa	LLMs *encoder* x *decoder* em tarefas de classificação e NER.

Fonte: De autoria própria.

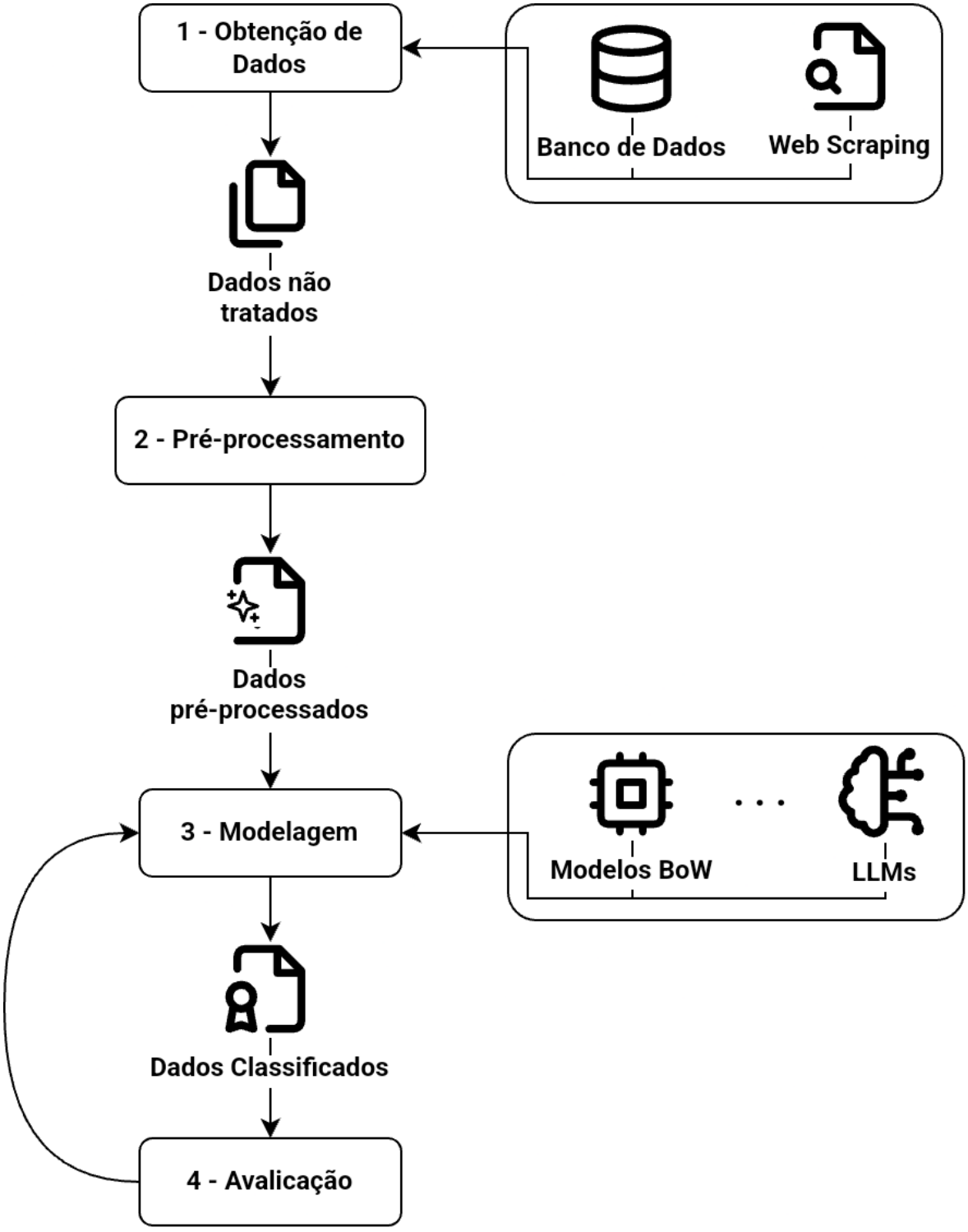
Capítulo 4

Metodologia

Neste capítulo, são apresentadas as definições da metodologia aplicada nesta pesquisa, que foi baseada na CRISP-DM (SHEARER, 2000), contemplando quatro etapas: coleta de dados, pré-processamento, modelagem e avaliação. A etapa de coleta de dados visa obter dados do DOU para posterior processamento. Na etapa de pré-processamento, são realizados os processos de padronização dos dados de forma a assegurar conformidade com o formato esperado, além da remoção de elementos textuais indesejados, como tags HTML e caracteres codificados. Quanto à etapa de modelagem, a mesma consiste na utilização de modelos de aprendizagem de forma a atingir a tarefa-alvo, que consiste na classificação de publicações do DOU quanto ao domínio tributário, por meio do treinamento a partir do conjunto de dados obtido. Por fim, na última etapa, realiza-se a avaliação dos modelos treinados aferindo, através de métricas apropriadas, a eficácia e eficiência dos mesmos. Na Figura 4.1 é ilustrado o fluxo descrito.

Cada uma das etapas do fluxo de processamento oferece intercambiabilidade de soluções, ou seja, permite que sejam utilizadas diferentes soluções, desde que solucionem o mesmo problema-alvo de cada etapa. Por exemplo, a etapa de obtenção de dados pode ser realizada tanto com a utilização de um *Web Scraper* como por meio de um acesso direto a um banco de dados, utilizando assim um conjunto já existente de dados. Essa versatilidade também se estende para a etapa de modelagem, dado que vários algoritmos de AM e modelos de classificação podem ser empregados, a exemplo dos LLMs baseados em *transformers*. A etapa de pré-processamento também é passível dessa intercambiabilidade, dado que soluções de classificação diferentes podem possuir requisitos específicos e distintos.

Figura 4.1: Etapas da metodologia descrita para atender ao problema de classificação de publicações tributárias do Diário Oficial da União.



Fonte: De autoria própria.

O conteúdo deste capítulo é dividido de acordo com as etapas metodológicas descritas, obedecendo o seguinte formato: na Seção 4.1 é descrita a etapa de coleta de dados; na Seção 4.2 são detalhados os pré-processamentos realizados no *corpus* utilizado; na Seção 4.3 estão listados os modelos de aprendizagem selecionados para realização de experimentos; já na Seção 4.4, é discutida a abordagem adotada na avaliação dos resultados; por fim, na Seção 4.5, são contempladas as considerações finais acerca das definições metodológicas desta pesquisa.

4.1 Coleta de Dados

Para classificar atos tributários no Diário Oficial da União foi utilizada a abordagem de classificação por aprendizagem supervisionada. Sendo assim, foi necessário construir um *corpus* composto de publicações do DOU. Foi utilizado como ponto de partida para construção do *corpus* um conjunto de dados anotados previamente, entretanto, composto apenas por publicações tributárias.

Foi então realizada uma análise descritiva do conteúdo deste conjunto de dados já existente, que será detalhada a seguir. A partir daí, implementou-se um *Web Scraper* com o objetivo de obter publicações do DOU. Por fim, foi empregada uma estratégia de correspondência, do inglês, *matching*, de forma a identificar as publicações existentes no conjunto de dados previamente anotado dentre as publicações obtidas diretamente do DOU. Dessa forma, buscando distinguir as instâncias positivas das negativas, ou seja, tributárias e não tributárias, respectivamente. Ao final deste processo, resultou-se um *corpus* do DOU anotado para a tarefa de classificação binária de texto no domínio tributário.

Nas subseções a seguir serão detalhadas cada uma das etapas descritas anteriormente, empregadas de forma a construir o *corpus*. Além disso, também serão discutidos os impactos do cenário altamente desbalanceado.

4.1.1 Conjunto de Dados Previamente Anotado

Este conjunto de dados, referido doravante como 'conjunto de dados original', foi manualmente anotado por uma empresa privada brasileira com interesse no monitoramento de alterações tributárias advindas dos Diários Oficiais. Cada registro presente no conjunto de dados

original possui 27 atributos, dos quais foram extraídos apenas os atributos de interesse: entidade de publicação (por exemplo, o DOU), título da publicação, data da publicação e texto completo. Um exemplo do conjunto de dados original é apresentado na Tabela 4.1.

A empresa responsável pela criação do conjunto de dados original realiza o monitoramento do conteúdo publicado no DOU, notificando, em relatórios, publicações no domínio tributário. Porém, este serviço, além de possuir um custo atrelado, possui um atraso de pelo menos três dias na notificação da existência de conteúdo tributário a partir de sua publicação. Tal atraso na notificação pode impactar empresas que lidam com dados tributários, atrasando potenciais ações que realizam a conformidade de seus sistemas com a norma tributária vigente. Esse ponto é de essencial importância, sobretudo em soluções que possuem abrangência em todo o território nacional.

Dado que quaisquer alterações na legislação tributária podem gerar impactos significativos, o monitoramento constante de tais alterações torna-se vital para o sucesso do negócio. Em virtude da origem privada deste conjunto de dados, não é permitida a disponibilização pública do *corpus* gerado nesta pesquisa, sem consentimento expresso da empresa detentora do conjunto de dados original.

O conjunto de dados original é composto de 54.466 publicações tributárias, tendo a mais antiga sido publicada no ano de 1964. Além disso, estão compreendidas as publicações de 130 entidades, dentre elas o Diário Oficial da União, dos Estados, Distrito Federal e diversos Municípios. Aproximadamente 97,41% dos dados possuem data de publicação posterior ao ano de 2010, conforme ilustrado na Figura 4.2. As publicações do DOU representam, aproximadamente, 20,26% de todo o conteúdo existente no conjunto de dados original, totalizando 11.858 registros e sendo o órgão publicador mais relevante, conforme exibido na Figura 4.3.

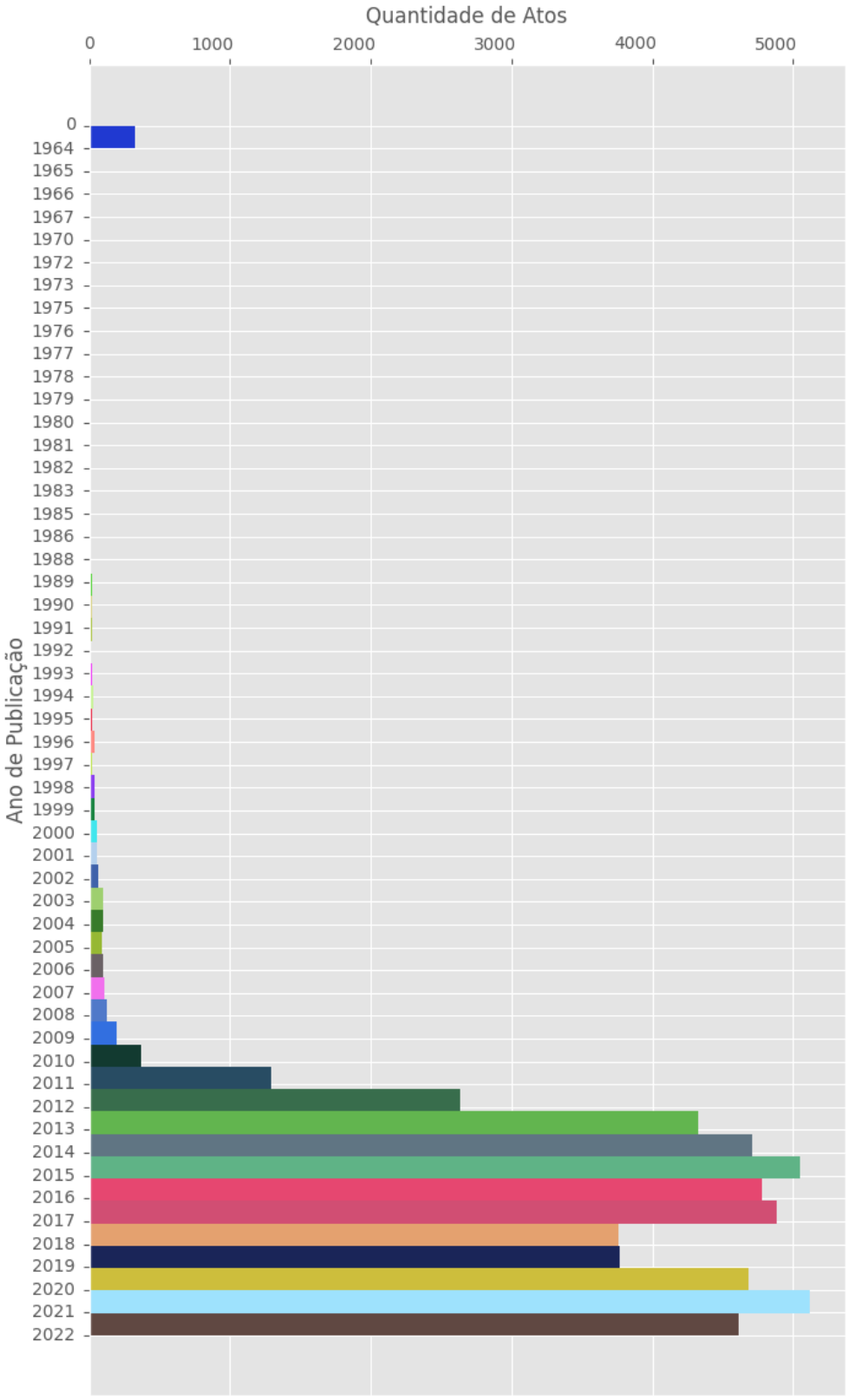
Dado o escopo da investigação realizada neste trabalho, o conjunto de dados original foi refinado e filtrado de forma a serem utilizados apenas os dados referentes ao DOU. Dessa forma, nesta pesquisa é endereçado o Diário Oficial de maior relevância no Brasil, dada sua influência nacional, além de ser o que possui mais registros tributários anotados dentre os demais Diários Oficiais existentes no conjunto de dados original, conforme evidenciado pelo gráfico da Figura 4.3. Quando analisam-se apenas as publicações do DOU percebe-se, vide Figura 4.4, que a distribuição e frequência das publicações do DOU, ao longo dos anos,

Tabela 4.1: Fragmentos de publicações do DOU extraídas de exemplos do conjunto de dados original.

Fragmento do Texto de Publicação do Conjunto Original - Publicações Tributárias do DOU	
Título:	Ato COTEPE nº 10/2018
Data de Publicação:	29/05/2018
Texto completo:	"Preço Médio Ponderado ao Consumidor Final (PMPF) de combustíveis. O Secretário-Executivo do Conselho Nacional de Política Fazendária - CONFAZ, no uso das atribuições que lhe são conferidas pelo inciso IX, do art. 5º do Regimento do CONFAZ [...], divulga que os Estados e o Distrito Federal adotarão, a partir de 1º de junho de 2018, o seguinte preço médio ponderado ao consumidor final (PMPF) para os combustíveis referidos [...]"
Título:	Instrução Normativa RFB nº 2.002/2020
Data de Publicação:	31/12/2020
Texto completo:	"Altera a Instrução Normativa RFB nº 680, de 2 de outubro de 2006 , que disciplina o despacho aduaneiro de importação. O Secretário Especial da Receita Federal do Brasil, no uso da atribuição que lhe confere o inciso III do art. 350 do Regimento Interno da Secretaria Especial da Receita Federal do Brasil, aprovado pela Portaria [..], no Decreto nº 660, de 25 de setembro de 1992 , nos arts. 542 a 579-A do Decreto nº 6.759, de 5 de fevereiro de 2009 - Regulamento Aduaneiro [...]"
Título:	Decreto nº 7.979/2013
Data de Publicação:	09/04/2013
Texto completo:	"Altera o Decreto nº 6.022, de 22 de janeiro de 2007, que instituiu o Sistema Público de Escrituração Digital - Sped. A PRESIDENTA DA REPÚBLICA, no uso da atribuição que lhe confere o art. 84, caput, inciso IV, da Constituição,DECRETA: Art. 1º O Decreto nº 6.022, de 22 de janeiro de 2007, passa a vigorar com as seguintes alterações: "Art. 2º O Sped é instrumento que unifica as atividades de recepção, validação, armazenamento e autenticação de livros e documentos que integram a escrituração contábil e fiscal dos empresários e das pessoas jurídicas, inclusive imunes ou isentas, mediante fluxo único, computadorizado, de informações [...]"
Título:	Protocolo ICMS nº 87/2014
Data de Publicação:	11/12/2014
Texto completo:	Altera o Protocolo ICMS 193/09, que dispõe sobre a substituição tributária nas operações com ferramentas. Os Estados de Minas Gerais, Paraná, Rio de Janeiro, Rio Grande do Sul e Santa Catarina, neste ato representados pelos seus respectivos Secretários de Fazenda, considerando o disposto nos arts. [...], resolvem celebrar o seguinte PROTOCOLO Cláusula primeira A cláusula primeira do Protocolo ICMS 193/09, de 11 de dezembro de 2009, passa a vigorar com a seguinte redação: [...]"

Fonte: De autoria própria.

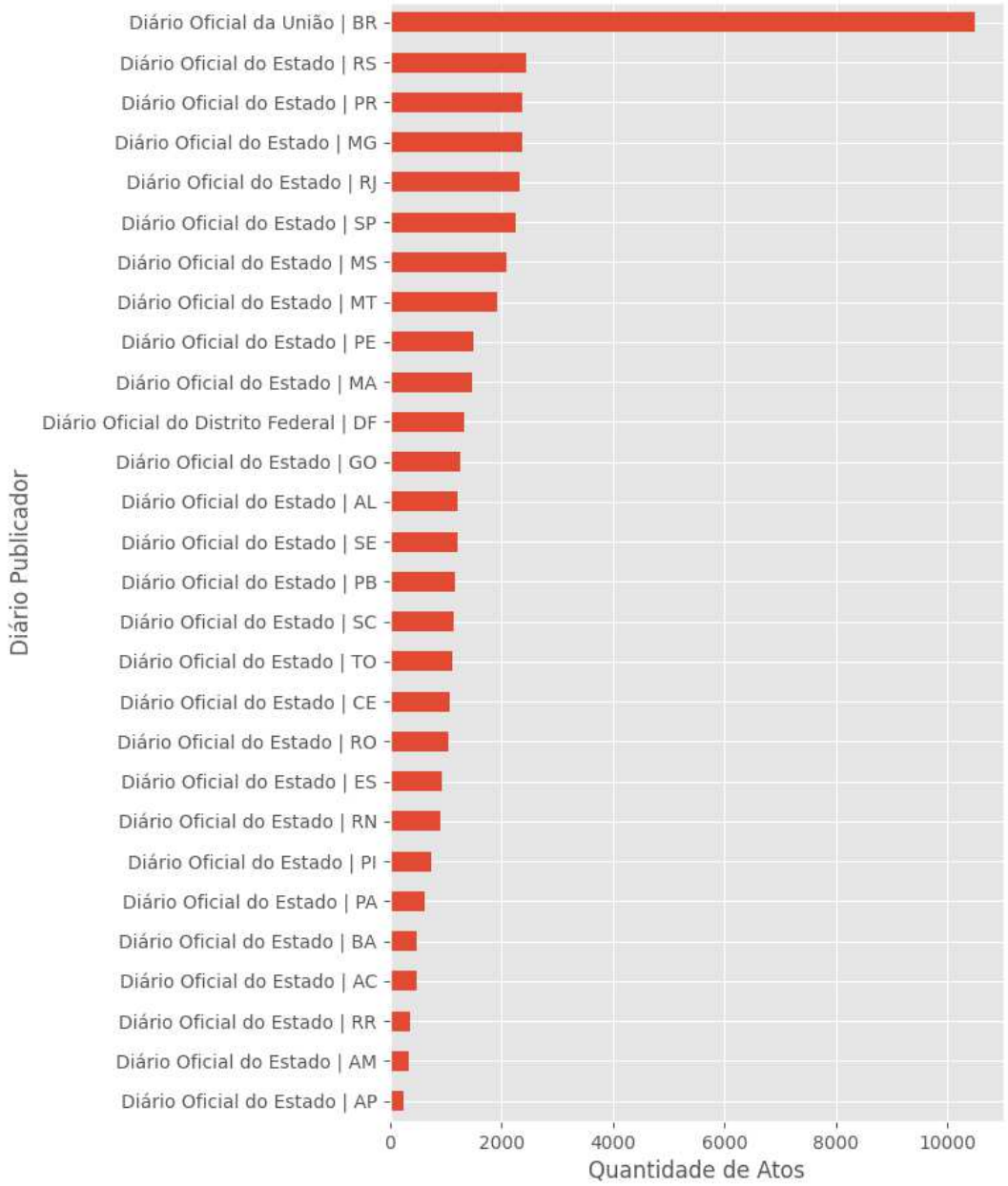
Figura 4.2: Gráfico de Histograma descrevendo a distribuição e frequência da data de publicação dos atos existentes no conjunto de dados original completo (cores adicionadas para auxiliar na diferenciação entre classes, não possuindo significado).



Fonte: De autoria própria.

apresenta comportamento semelhante ao observado no conjunto completo dos dados.

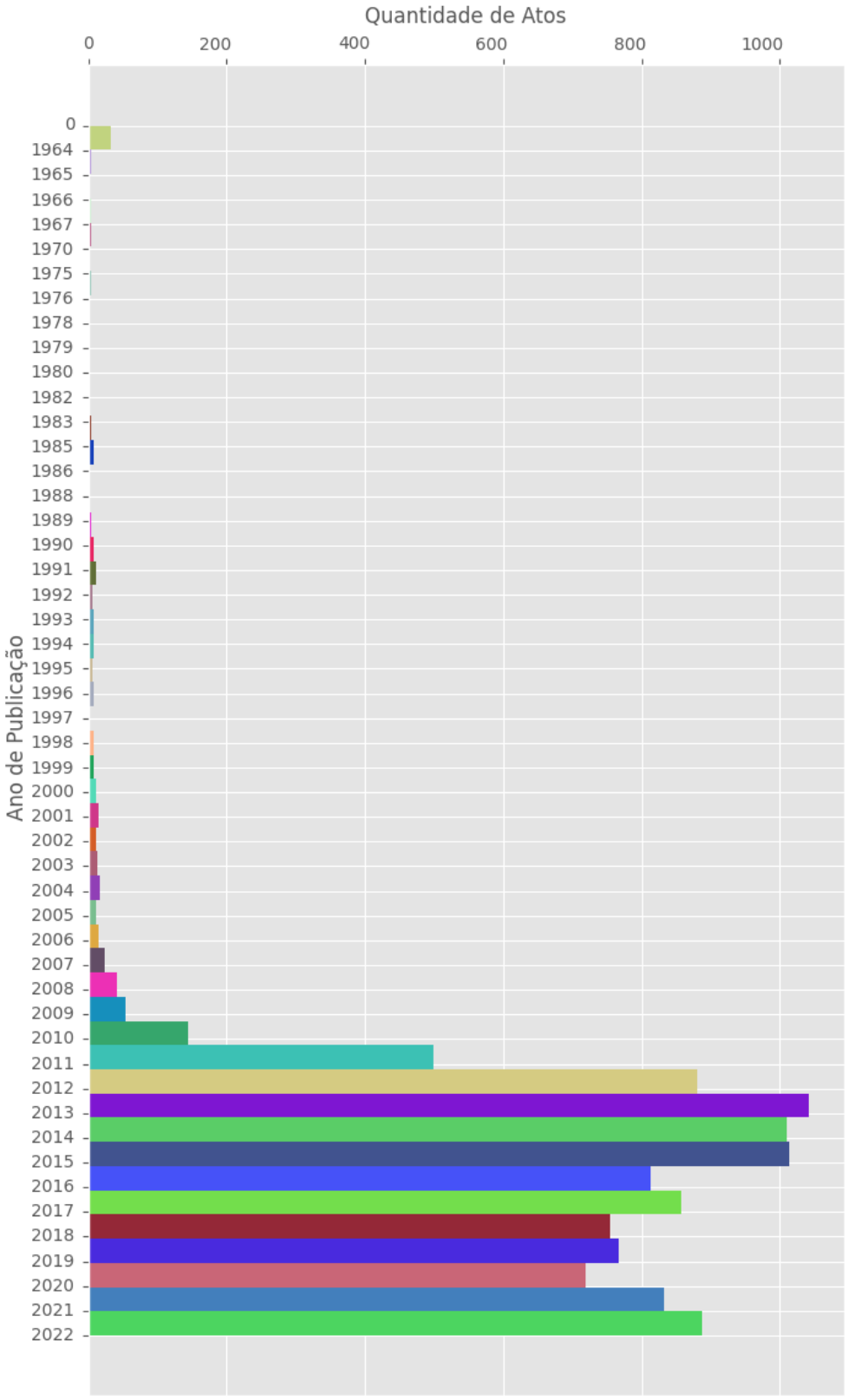
Figura 4.3: Gráfico de Barras descrevendo a quantidade de publicações existentes no conjunto de dados original para os Diários Oficiais dos Estados, Distrito Federal e União.



Fonte: De autoria própria.

Ao observar os gráficos de histograma presentes nas Figuras 4.2 e 4.4 é possível verificar a existência de publicações na classe 0-1964, primeira no eixo dos anos de publicação. Estes registros não possuem data de publicação registrada e foram removidos dos dados utilizados nas etapas seguintes de processamento. De forma a também verificar o comportamento da

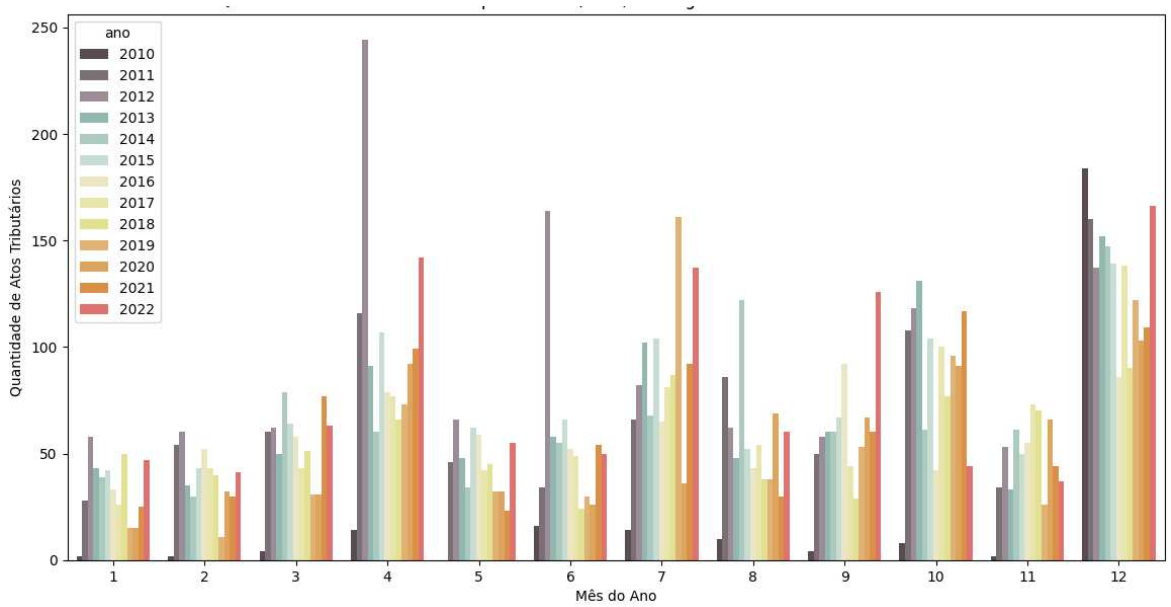
Figura 4.4: Gráfico de Histograma descrevendo a distribuição e frequência do ano de publicação dos atos existentes no conjunto de dados original filtrado pelo Diário Oficial da União (cores adicionadas para auxiliar na diferenciação entre classes, não possuindo significado).



Fonte: De autoria própria.

distribuição ao longo dos meses do ano, foi realizada uma análise da frequência de ocorrência de publicações tributárias observada por mês e por ano de ocorrência, conforme exibido na Figura 4.5. Para essa análise, foi considerado apenas o período compreendido de 2010 a 2022, dado ser o período contendo a maior fatia de dados.

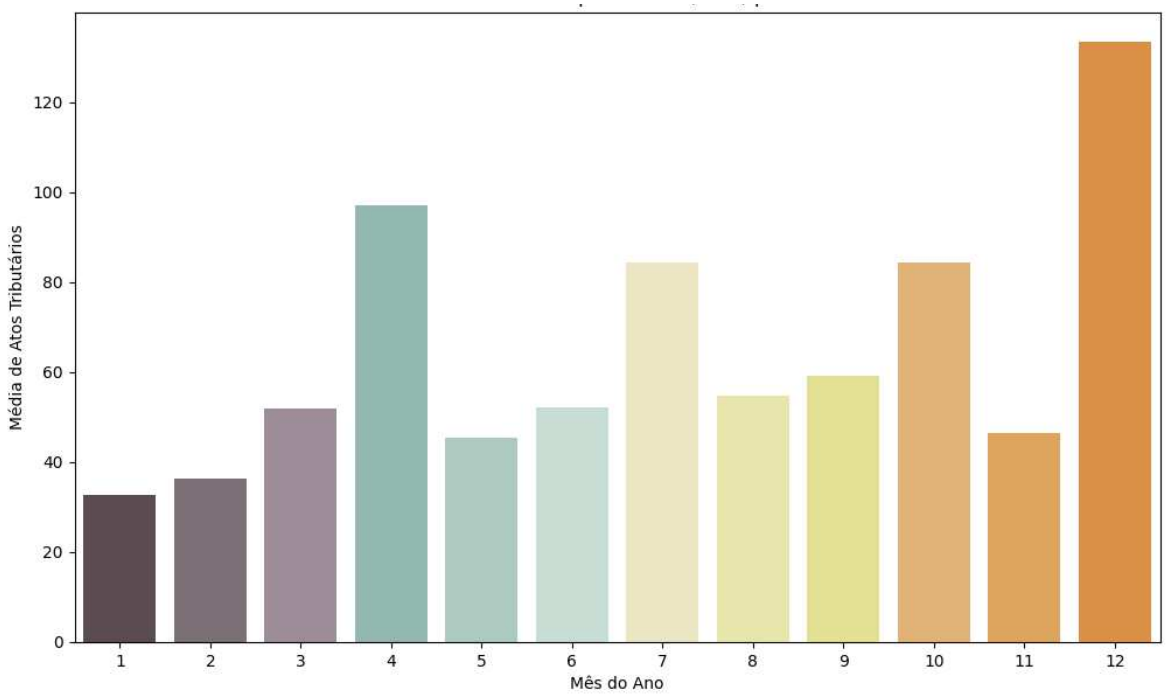
Figura 4.5: Gráfico de Barras da Quantidade de Atos Tributários por Mês ao longos dos anos de 2010 a 2022 para dados do Diário Oficial da União.



Fonte: De autoria própria.

Observou-se, pois, um padrão na quantidade de atos publicados ao longo do ano, onde os meses de abril, julho, outubro e dezembro apresentaram uma maior ocorrência quando comparados aos demais meses no escopo de num mesmo ano. De forma a verificar essa observação de forma mais clara, foi calculada a média da quantidade de atos tributários publicados em cada um dos meses ao longo dos anos de 2010 a 2022, conforme pode ser observado na Figura 4.6. Ficou então evidenciada a mesma conclusão tirada a partir do gráfico anterior. Dentre os meses citados, o mês de dezembro apresentou a maior frequência média de ocorrência de publicações dentre o período analisado.

Figura 4.6: Gráfico de Barras da Quantidade Média de Atos Tributários por Mês entre os anos de 2010 e 2022 para dados do Diário Oficial da União.



Fonte: De autoria própria.

4.1.2 Obtenção de Conteúdo do Diário Oficial da União

O conjunto de dados original consiste apenas de instâncias positivas associadas ao problema de classificação de publicações tributárias. Diante disso, foi criado um conjunto de dados automaticamente anotado, compreendendo tanto instâncias positivas quanto negativas. As publicações do DOU podem ser acessadas diretamente por meio da sua página da Web[1], sendo subdividido em três principais jornais. O primeiro é relacionado a atos normativos, o segundo a atos de pessoal e o terceiro é referente a contratos, editais e avisos. De forma a abranger a maior quantidade de domínios diferentes, todos os três jornais foram englobados no *corpus* criado.

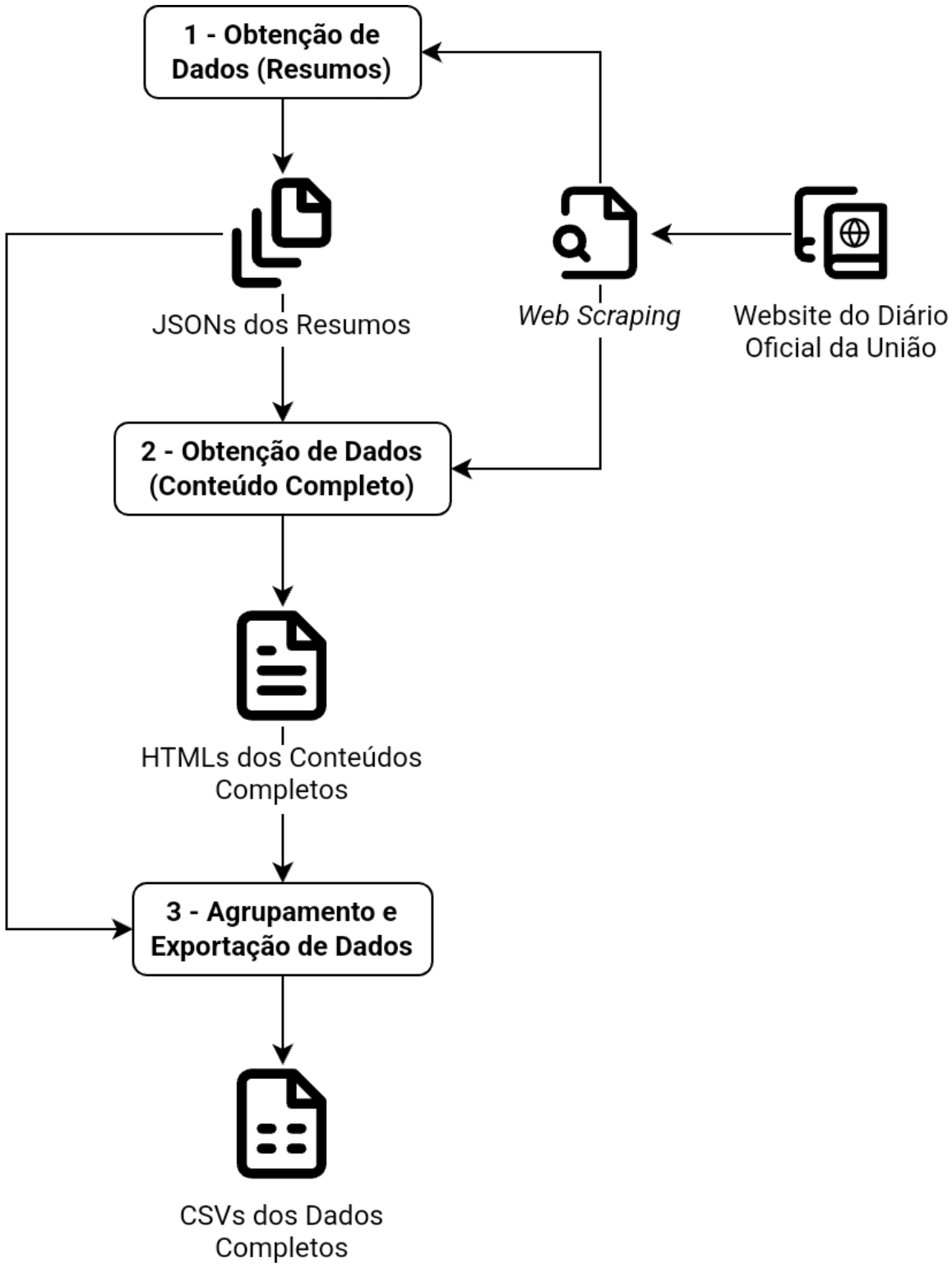
Para realizar a coleta dos dados disponíveis a partir do DOU, foi desenvolvido, em Python versão 3.9.16, um *Web Scraper* específico para a página do respectivo Diário Oficial, onde foram obtidos, inicialmente, dados de publicações realizadas desde 2013, sendo este o primeiro ano com dados disponíveis para consulta prévia, até 2022, ano final do levantamento realizado nesta pesquisa. Dadas as características específicas existentes na página da Web que disponibiliza as publicações do DOU, fez-se necessário segmentar a obtenção e o processamento do conteúdo do Diário Oficial em três etapas. A organização dessas etapas e seus relacionamentos são exibidos na Figura 4.7.

A etapa inicial de processamento do *Web Scraper* consiste em obter as informações resumidas de cada publicação disponível na página, para cada dia do período de processamento especificado, de 2013 a 2022, nos três jornais existentes. É possível acessar a lista de todas as publicações realizadas para um determinado dia e jornal utilizando parâmetros de URL. O link de acesso para para o DOU é “https://www.in.gov.br/leiturajornal”, podendo ser acompanhado dos parâmetros “data”, em formato “DD-MM-YYYY”, e “secao”, recebendo os valores “do1”, “do2” ou “do3”, respectivo a cada um dos jornais já descritos anteriormente. Assim, para acessar as publicações realizadas no jornal de atos normativos do dia 07/06/2022, o link completo é “https://www.in.gov.br/leiturajornal?data=07-06-2022&secao=do1”.

A página exibida, dada uma combinação de parâmetros válida, lista todas as publicações existentes para a respectiva combinação de parâmetros, conforme demonstrado na Figura 4.8. Os dados utilizados para exibição do conteúdo na página são encaminhados juntamente do conteúdo HTML, no formato JSON sob uma tag “<script>”, responsável em HTML por

[1] https://www.in.gov.br/leiturajornal

Figura 4.7: Etapas de processamento do *Web Scraper* de obtenção de dados do Diário Oficial da União.



Fonte: De autoria própria.

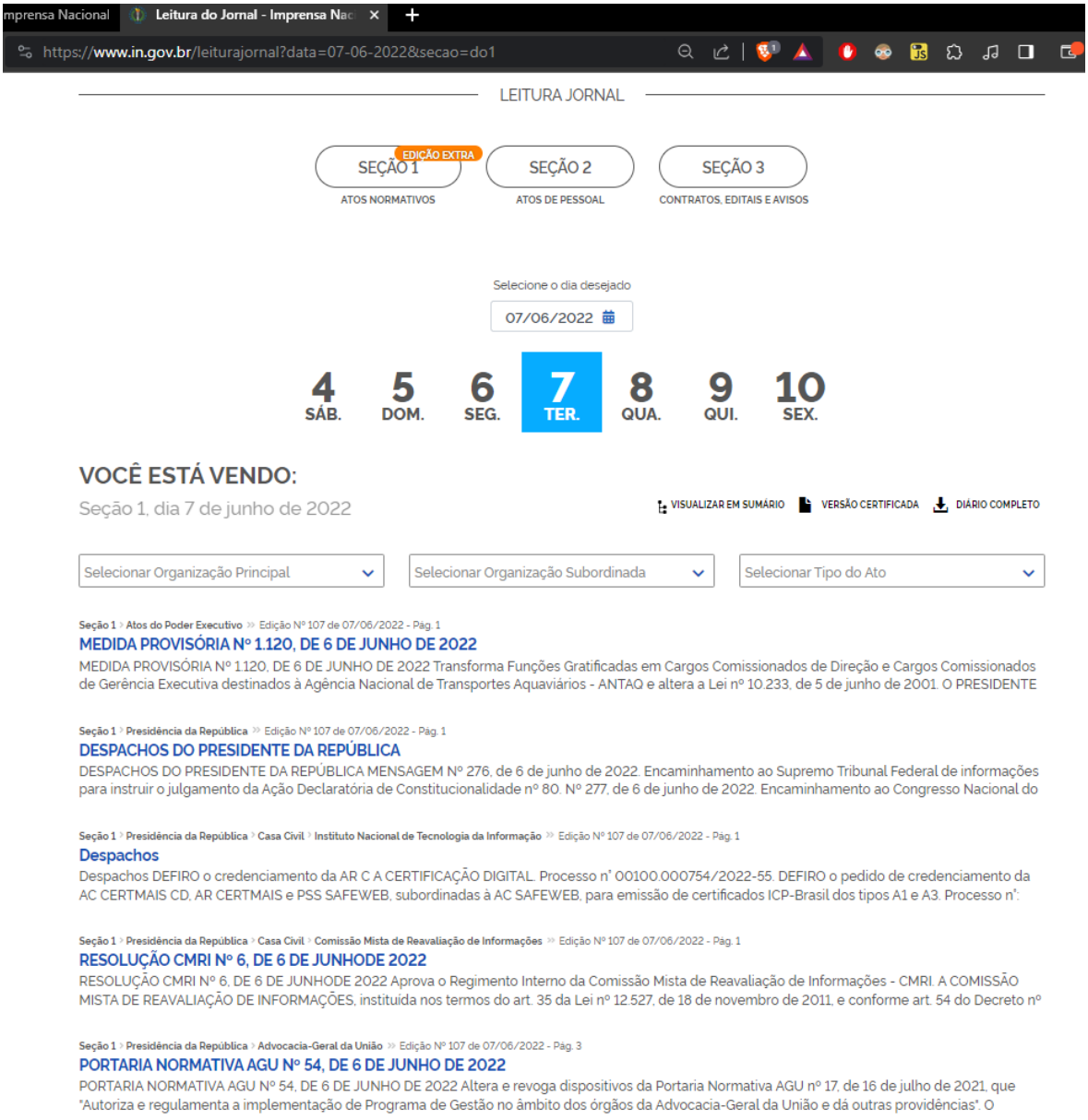
armazenar trechos de código em JavaScript, contendo o parâmetro ‘id="params"’. Foi então utilizado este seletor para a obtenção do conteúdo JSON por parte do *Web Scraper*.

O conteúdo JSON presente na requisições HTTP à página do DOU contém as seguintes informações de cada publicação: título, data de publicação, número da edição do Diário Oficial, número da página na edição, tipo de publicação, órgãos públicos associados, conteúdo resumido da publicação, ou ementa, com limitação de 403 caracteres e, por fim, fragmento de URL para acessar o conteúdo completo da publicação.

Após a obtenção dos arquivos JSON contendo todas as publicações realizadas entre 2013 e 2022, iniciou-se a segunda etapa do processamento que teve foco na obtenção do conteúdo completo das publicações. O conteúdo completo de cada publicação, também acessível via clique no *hiperlink* presente em cada um dos itens da listagem exibida na Figura 4.8, foi acessado utilizando o fragmento de URL obtido a partir do JSON obtido na etapa anterior. A URL base para acessar cada publicação completa foi a “https://www.in.gov.br/web/dou/-/”, que concatenada com cada fragmento presente no JSON do resumo das publicações, levou ao conteúdo completo. Tomando como exemplo um dos fragmentos de URL obtidos a partir do JSON de publicações resumidas, “medida-provisoria-n-1.120-de-6-de-junho-de-2022-405908250”, o link para acessar o conteúdo completo é: “<https://www.in.gov.br/web/dou/-/medida-provisoria-n-1.120-de-6-de-junho-de-2022-405908250>”. Assim, cada fragmento de URL, quando concatenado com o link base, levou ao website do DOU, em uma página contendo o conteúdo completo da publicação.

De forma a garantir que todas as publicações de cada dia foram obtidas com sucesso, cada registro obtido a partir dos JSONs salvos na primeira etapa da execução recebeu um identificador único (UUID). Todo o conteúdo obtido foi organizado em sistema de arquivos, onde cada data possui seu diretório contendo arquivos HTML cujo nome é o próprio identificador da publicação obtida. Ao final do processamento de cada dia, todos os identificadores das publicações são percorridos e é verificada a existência do arquivo HTML de mesmo nome, em caso de ausência do arquivo, significa que o conteúdo daquele respectivo ato não foi obtido com sucesso. Por fim, a terceira etapa consiste no processamento de todo o conteúdo completo obtido e organizado em sistema de arquivos na segunda etapa, unindo-o ao conteúdo resumido das publicações obtido na primeira etapa, a partir do JSON enviado em conjunto com o conteúdo HTML. O resultado final consiste em arquivos CSV, separados

Figura 4.8: Captura de tela da página do Diário Oficial da União exibindo publicações no formato de listagem.



Fonte: Website do Diário Oficial da União.

por ano de publicação de forma a diminuir o tamanho dos arquivos resultantes, contendo todo o *corpus*.

4.1.3 Identificação de Correspondência de Publicações no Conjunto de Dados

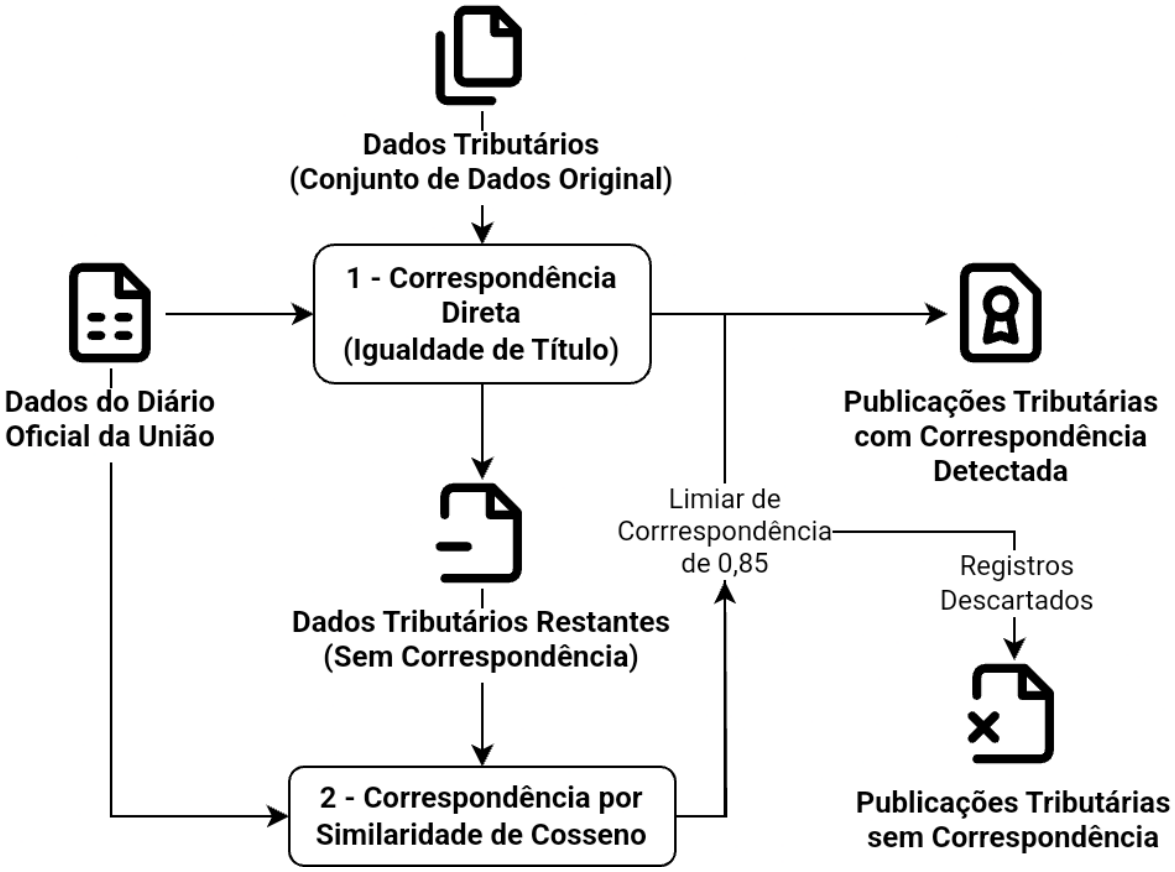
A anotação manual de conjuntos de dados é, normalmente, uma tarefa extremamente custosa. Para o escopo do problema-alvo deste trabalho, foi possível valer-se da existência do conjunto de dados original, onde foram empregadas, a partir daí, estratégias para a obtenção de um conjunto de dados anotado automaticamente. Dessa forma, obtendo um conjunto de dados passível de utilização em experimentos de classificação de atos extraídos do DOU com foco no domínio tributário.

Sendo assim, a estratégia base utilizada para a anotação automática do novo conjunto de dados foi utilizar os títulos das publicações como base para correspondência dos registros entre os conjuntos de dados. Foram utilizados tanto o conjunto de dados original como o *corpus* obtido de atos publicados no DOU utilizando o *Web Scraper* fruto deste trabalho. Todas as publicações existentes no conjunto de dados original estão compreendidas dentre as extraídas diretamente do DOU, dado que ambas possuem a mesma origem. Com base nessa informação, há garantia de que todas as publicações presentes no conjunto de dados original possuam uma correspondência.

O título das publicações do DOU é altamente informativo, abrangendo o tipo de publicação, o nome, acrônimo ou sigla de uma entidade pública e um identificador exclusivo para aquele ano. Por exemplo, o título "Portaria FBN Nº 61" designa a Portaria Nº 61 publicada pela Fundação Biblioteca Nacional (FBN). Isso torna cada título um identificador único de cada publicação, colaborando na justificativa da abordagem escolhida de correspondência a partir dos títulos de publicações.

Foi empregada a seguinte abordagem para identificar correspondências entre ambos os conjuntos de dados: verificou-se primeiro a existência de títulos iguais, onde, caso não encontrados, foi realizado o cálculo da similaridade de cosseno entre os títulos. A escolha da utilização da similaridade de cosseno deu-se frente a ser uma das estratégias mais utilizadas em tarefas de recuperação da informação (MANNING; RAGHAVAN; SCHUTZE, 2008). O

Figura 4.9: Fluxo da estratégia de identificação de correspondência entre títulos de publicações de forma a anotar o conjunto de dados automaticamente.



Fonte: De autoria própria.

fluxo completo da estratégia de correspondência encontra-se resumido na Figura 4.9. Cada publicação tributária passou por uma tentativa de correspondência com todas as publicações extraídas do DOU na mesma data de publicação.

De forma a contribuir com a explanação da abordagem selecionada, segue um exemplo prático considerando uma data aleatoriamente selecionada. Para a data em questão, todos os títulos inerentes ao conjunto de dados original e aqueles diretamente extraídos do DOU foram selecionados para análise. Essa análise consistiu em buscar os títulos existentes no conjunto de dados original dentre os dados recentemente extraídos do DOU. O processo foi iniciado com uma avaliação de igualdade direta entre strings, sendo então apontados os títulos idênticos. Em seguida, para cada um dos títulos de publicações tributárias ainda pendentes de correspondência, foram calculados os valores da similaridade de cosseno dentre o

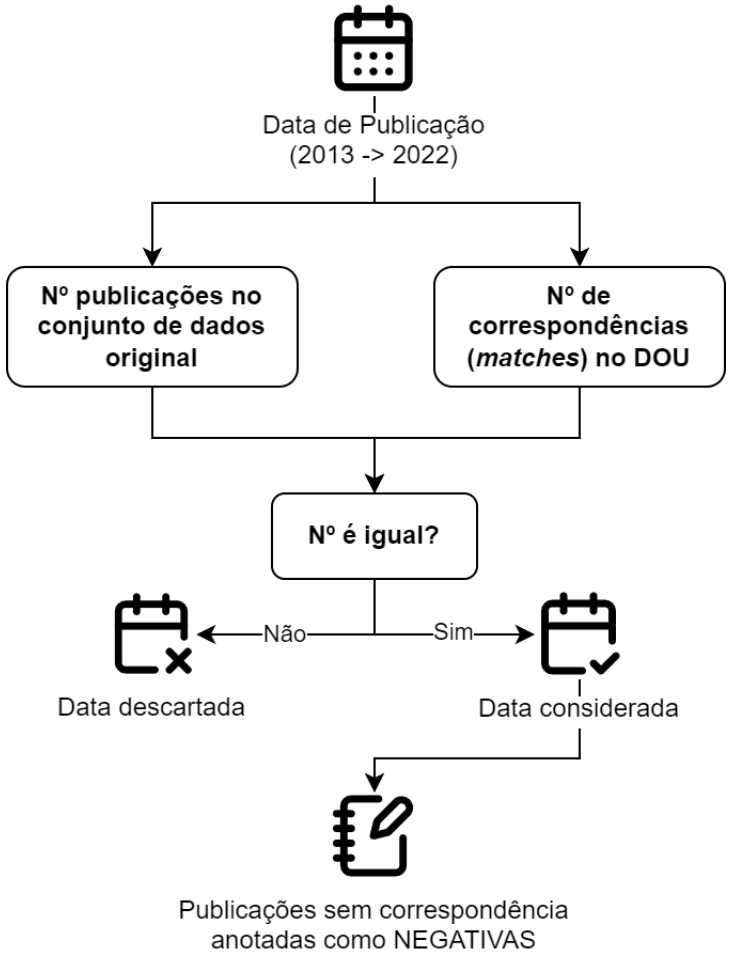
título selecionado frente aos demais existentes nos dados obtidos do DOU, desde que ainda não haja correspondência. A preferência por este último método de identificação de correspondência parte do reconhecimento de que os títulos de publicações constituem expressões textuais concisas, seguindo o padrão descrito anteriormente. Dessa forma, mitigando possíveis discrepâncias nos estilos de escrita que possam ter se manifestado durante a formulação do conjunto de dados original.

De forma a evitar erros decorrentes de métricas de similaridade elevadas entre títulos de publicações distintas, foi estabelecido um limiar mínimo de similaridade cosseno de 0,85. O valor utilizado como limiar mínimo foi determinado empiricamente por meio de experimentação. Assim, só foram levados em conta registros que apresentaram uma métrica de similaridade superior ao limite estipulado. O título com a métrica de similaridade mais alta foi considerado como sendo a correspondência. Ao aplicar o passo a passo descrito no conjunto de dados completo, 1.206 (10,17% do conjunto original) publicações tributárias foram identificadas com títulos idênticos frente a 3.476.943 publicações extraídas do DOU. Com relação à estratégia utilizando similaridade de cosseno, foram automaticamente identificadas outras 2.723 (22,96% do conjunto original) publicações. Ao final, foram identificadas 3.929 correspondências de publicações tributárias entre conjuntos de dados, representando um mapeamento de 33,17% do conjunto original para o novo *corpus* obtido diretamente do DOU.

O passo seguinte consistiu na anotação automática das amostras negativas, seguindo conforme o fluxo disposto na Figura 4.10. Para tal, foi levado em conta possíveis erros de rotulagem, onde publicações de domínio tributário poderiam ser erroneamente rotuladas como não tributárias se todas as demais publicações que não tiveram correspondência identificada fossem designadas como amostras negativas. Sendo assim, para aumentar a confiança no conjunto de dados, foram selecionadas as datas em que as publicações tributárias foram identificadas na íntegra, ou seja, removendo as datas cujo número de correspondências diferiu do total de publicações tributárias observado no conjunto de dados original. Para as datas selecionadas, todas as publicações no *corpus* obtido a partir do DOU sem correspondência foram anotadas como amostras negativas.

Exemplificando a anotação automática das amostras negativas, suponhamos um cenário fictício no qual constam 10 publicações tributárias, no conjunto de dados original, e um total

Figura 4.10: Fluxo da estratégia de Anotação Automática de Amostras Negativas.



Fonte: De autoria própria.

de 1000 publicações foram obtidas diretamente do DOU pelo *Web Scraper*. Após aplicar a estratégia de correspondência descrita anteriormente, suponhamos que todos os 10 registros presentes no conjunto de dados original foram encontrados nas 1000 publicações do DOU existentes para uma determinada data. Dado que todos os registros positivos foram encotrandos para a referida data, os 990 registros restantes serão anotados como negativos. Considerando um outro cenário fictício, em uma outra data, sendo que neste, ao invés da estratégia identificar todos os 10 registros tributários em meio aos atos do DOU, apenas 6 foram encontrados, ainda existindo outros 4 atos tributários em meio aos 994 sem correspondência. Dado que não foi possível determinar quais os atos tributários restantes, anotar todas as demais publicações como negativas implicaria em anotar 4 registros positivos como negativos, impactando negativamente os resultados. Sendo assim, esta data seria descartada,

em conjunto com todos os registros associados à mesma, sejam positivos ou negativos.

A metodologia escolhida para a anotação de amostras negativas tem implicações no tamanho do conjunto de dados, reduzindo o número de instâncias positivas ao desconsiderar registros de datas que possuem correspondência incompleta. No entanto, ao desconsiderar até mesmo as instâncias positivas identificadas, a proporção original dos dados é preservada, mantendo fidelidade ao observado no cenário real e melhorando assim a qualidade geral do conjunto de dados obtido. Como resultado, o conjunto de dados final a ser utilizado na etapa de experimentação foi composto por 1.444 publicações tributárias e 1.640.532 publicações não tributárias, todas oriundas do DOU. É importante ressaltar a distribuição altamente desbalanceada deste conjunto de dados, onde as publicações tributárias representam aproximadamente 0,09% do total.

4.2 Pré-processamento

Anteriormente à utilização do *corpus* construído, foram realizadas diversas etapas de pré-processamento no conteúdo das publicações, a saber: remoção de tags HTML, decodificação de texto HTML codificado, conversão de todo o texto para apenas letras minúsculas, substituição de quebras de linha por espaços e remoção de palavras irrelevantes (*stopwords*), sendo este último apenas para o conjunto aplicado aos modelos do tipo *Bag-of-Words* (BoW).

Além disso, o *corpus* foi subdividido em conjuntos de treinamento, validação e teste, tomando como base as datas de publicação para a fragmentação. Na Tabela 4.2 é descrita a distribuição dos dados no *corpus*. Para abordar o alto desbalanceamento do conjunto de dados utilizado, onde 99% dos dados correspondem às instâncias negativas, e comparar os resultados de diferentes estratégias de balanceamento do conjunto de treinamento, foram construídos dois conjuntos de treinamento: um balanceado e outro levemente desbalanceado. Para o conjunto de dados balanceado, foi realizada a seleção aleatória de um registro negativo para cada positivo existente, obtendo assim um balanceamento a partir da classe minoritária (*undersampling*). Já para o conjunto de dados desbalanceado, foi realizada a seleção aleatória de três registros negativos a partir de cada instância positiva. O conjunto de treinamento compreendeu dados de 2013 a 2021.

Caso fosse mantida a proporção real dos dados para o conjunto de treinamento, seria ne-

Tabela 4.2: Particionamento de treinamento, validação e teste do conjunto de dados utilizado.

Conjunto de Dados	Nº de Positivos	Nº de Negativos	Proporção	Período
Treinamento 1 (Balanceado)	1.093	1.093	1:1	2013 - 2021
Treinamento 2 (Desbalanceado)	1.093	3.279	1:3	2013 - 2021
Validação	36	39.348	1:1.093	2021
Teste	315	344.254	1:1.093	2022

Fonte: De autoria própria.

cessário adicionar 1.196.835 amostras negativas. A inclusão de tamanha quantidade de dados pode prolongar significativamente os tempos de treinamento dos modelos, além de possivelmente direcionar modelos a exibirem viés direcionado à classe majoritária, ofuscando assim a classe minoritária. Trabalhos recentes dedicaram esforços no tratamento de dados desbalanceados em classificação de texto (YANG *et al.*, 2022; HASIB *et al.*, 2023), com soluções seguindo na mesma linha da aplicada neste trabalho, inserindo desbalanceamento no conjunto de treinamento. Os autores destacaram contextos desbalanceados que demonstraram uma melhoria significativa de desempenho a partir da inserção de um leve desbalanceamento nos dados de treinamento de modelos *transformer*, corroborando com a abordagem adotada nesta pesquisa.

O conjunto de validação, proveniente de dados de 2021, foi composto por 36 publicações tributárias e 39.348 publicações não tributárias. Para cada mês do ano, foram selecionados aleatoriamente três registros positivos e 3.279 registros negativos, excluindo-se aqueles já presentes no conjunto de treinamento, mantendo assim a proporção real do desbalanceamento do conjunto de dados. Optou-se por selecionar apenas os 36 registros positivos, para o conjunto de validação, por conta da pequena quantidade de registros positivos existentes no *corpus* como um todo, dado que uma maior quantidade de positivos no conjunto de validação implicaria em uma menor quantidade de positivos no conjunto de treinamento. O conjunto de teste abrangeu todas as publicações de 2022, contendo 315 instâncias positivas e 344.254 instâncias negativas.

O ano de publicação dos atos foi o principal critério para o particionamento dos dados, dada sua natureza temporal. Isso se deve à precedência das publicações, seguindo a ordem cronológica. Não respeitar essa ordem em dados temporais pode afetar significativamente a modelagem de problemas que possuem essa característica. Devido à baixa quantidade

de registros positivos, todo o intervalo temporal foi utilizado nos conjuntos de dados para os modelos, tanto no desbalanceado quanto no balanceado (ROCHA *et al.*, 2008; ROCHA *et al.*, 2013; SALLES *et al.*, 2015). Não foram realizadas análises adicionais sobre o efeito da variação da janela temporal nos resultados, pois isso está fora do escopo deste trabalho. Ao final, considerando a classe positiva, a proporção de *holdout* do conjunto de dados resultante foi de aproximadamente 78%/22% (Treino + Validação / Teste), com 3% do conjunto de treinamento destinado à validação.

Diferentemente dos LLMs de arquitetura *encoder*, os de arquitetura *decoder*, como o Llama 2, possuem como saída texto gerado a partir de um prompt passado como entrada, ao invés de uma probabilidade associada ao problema alvo. Para este problema em questão, ao invés dos valores "0" e "1", foi realizado um pré-processamento no conjunto de dados de forma a adaptá-lo para o experimento no formato de *prompt*. Os valores das saídas esperadas (valores anotados, ou *ground-truth*) foram convertidos para "positivo" e "negativo", simbolizando a ocorrência ou não de texto de cunho tributário. Além disso, cada elemento do conjunto de dados passou a ter quatro atributos de composição: contexto, instrução, entrada e saída. Um exemplo de instância do conjunto de dados neste formato pode ser observado na Tabela 4.3.

Dada a maior quantidade de dados de pré-treinamento do modelo utilizado ter sido em inglês, foi mantido este idioma para o texto presente no contexto, dado que é uma padronização na construção de *prompts* para treinamento de LLMs de arquitetura *decoder*. Já o restante do *prompt*, tanto a instrução como demais pontos, foi mantido em português.

4.3 Modelagem

De forma a alcançar o principal objetivo deste trabalho, o de obter um modelo capaz de classificar atos do DOU como tributários ou não tributários, diversos experimentos foram realizados. Utilizando o *corpus* automaticamente anotado, construído no escopo deste trabalho, um conjunto de modelos foram treinados para a tarefa alvo de classificação binária. Foi realizada uma análise comparativa compreendendo modelos tradicionais de classificação baseados em AM supervisionada, múltiplos LLMs com arquitetura baseada em *encoder*, e um modelo de LLM baseado na arquitetura *decoder*. Nesta seção, serão detalhados os modelos

Tabela 4.3: Exemplo de entrada para treinamento do modelo Llama 2.

Contexto	*Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.*
Instrução	Caso o texto a seguir seja de domínio tributário, contenha tributos, ou esteja relacionado a impostos, classifique como positivo. Caso contrário, como negativo.
Entrada	secretaria executiva ato declaratório nº 1, de 26 de janeiro de 2021 ratifica os convênios icms aprovados na 330ª reunião extraordinária do confaz, realizada no dia 21.01.2021 e publicados no dou em 22.01.21. o diretor da secretaria-executiva do conselho nacional de política fazendária - confaz, com fulcro no art. 5º da lei complementar nº 24, de 7 de janeiro de 1975, no uso das atribuições que lhe são conferidas pelo inciso x do art. 5° e pelo parágrafo único do art. 37 do regimento desse conselho, considerando a urgência requerida pelo secretários de estado da fazenda do amazonas; considerando que, após consulta realizada por meio do ofício circular sei nº 210/2021/me, as unidades federadas aprovaram, por unanimidade, a ratificação antecipada, declara ratificados os convênios icms a seguir identificados, celebrados na 330ª reunião extraordinária do confaz, realizada no dia 21 de janeiro de 2021: convênio icms 01/21 - revigora, dispõe sobre a adesão dos estados do amazonas, mato grosso do sul, pará, rio de janeiro e do distrito federal e altera o convênio icms 63/20, que autoriza as unidades federadas que menciona a conceder isenção do icms incidente nas operações e correspondentes prestações de serviço de transporte realizadas no âmbito das medidas de prevenção ao contágio e de enfrentamento à pandemia causada pelo novo agente do coronavírus (sars-cov-2); convênio icms 02/21 - autoriza as unidades federadas que menciona a conceder isenção do icms incidente nas operações e correspondentes prestações de serviço de transporte realizadas no âmbito das medidas de prevenção ao contágio e de enfrentamento à pandemia causada pelo novo agente do coronavírus (sars-cov-2); convênio icms 03/21 - autoriza as unidades federadas que menciona a conceder isenção do icms incidente nas saídas interestaduais, de oxigênio medicinal, destinadas ao estado do amazonas, em razão da crise sanitária provocada pelo covid-19 nas condições que especifica. renata larissa silvestre substituta
Saída	positivo

Fonte: De autoria própria.

selecionados, as configurações dos experimentos e as métricas utilizadas.

4.3.1 Modelos Tradicionais de Classificação

Para estabelecer resultados de referência, do inglês, *baselines*, usando métodos tradicionais de AM, os experimentos iniciais utilizaram duas abordagens: *Support Vector Machines* (SVM) e XGBoost. Para a vetorização de texto, foi adotado o TF-IDF para os modelos *baseline*, de forma a capturar os termos mais frequentes, endereçando o problema em um formato de abordagem simplificada, como uma busca direta de termos significativos. A escolha do

SVM deu-se por ser um modelo bastante utilizado na solução de problemas de classificação de texto (CERVANTES *et al.*, 2020; KAMRAN; SAEED; ALMAGHTHAWI, 2023; ABDALLA; AMER; RAVANA, 2023), sendo uma solução baseada em hiperplanos. O XGBoost foi escolhido por ser uma técnica baseada em árvore de decisão com aumento de gradiente e que também é bastante aplicado em tarefas de classificação (QI, 2020; CHEN *et al.*, 2022a; HENDRAWAN; UTAMI; HARTANTO, 2022).

Além dos modelos SVM e XGBoost selecionados para experimentação, de forma a obter uma solução *baseline* otimizada ao problema, foi utilizado o Auto-Sklearn (FEURER *et al.*, 2015; FEURER *et al.*, 2020), uma ferramenta de AM automatizada (do inglês, *automated machine learning*, ou AutoML). Seu objetivo consiste na seleção automática de algoritmos de AM, utilizando-se de otimização Bayesiana, meta-aprendizagem e *ensembles*, selecionando automaticamente modelos e hiperparâmetros de forma direcionada ao conjunto de dados utilizado. Dessa forma, a modelagem utilizando-se de técnicas tradicionais não limitou-se apenas aos modelos manualmente selecionados, apesar dos escolhidos serem bastante utilizados. Ao final, a partir da modelagem com objetivo na obtenção de *baselines*, é possível verificar o comportamento do problema diante de técnicas amplamente utilizadas em problemas de NLP e AM, fundamentando assim a discussão frente a modelos mais recentes e custosos, bem como os seus resultados.

4.3.2 Modelos de Linguagem Grandes (*Large Language Models* - LLMs) baseados em arquitetura *encoder*

Foram realizados experimentos utilizando LLMs baseados em arquitetura *encoder*. Os modelos selecionados incluíram uma diversidade de domínios base, dentre eles modelos multilíngues, incluindo BERT-multilingual e XLM-RoBERTa (CONNEAU *et al.*, 2019). Também foram realizados experimentos com modelos pré-treinados com foco em português, como BERTimbau, um modelo baseado no BERT, e ALBERTINA, modelo baseado em DeBERTa.

Ademais, também foram incorporados aos experimentos, modelos pré-treinados em domínios específicos, de forma a enriquecer a análise. Dentre os modelos adicionados, são exemplos o Legal-BERT (DOMINGUES, 2022) e BERTikal (POLO *et al.*, 2021), ambos pré-treinados utilizando um corpus de textos jurídicos. Também foi utilizado o modelo

Tabela 4.4: Resumo dos LLMs de arquitetura *encoder* utilizados.

Modelo	Nº de Parâmetros	Domínio	Idioma
BERTimbau-base	110M	Geral	Português
Legal-BERT	110M	Jurídico	Português
BERTikal	110M	Jurídico	Português
BERDOU	110M	DOU	Português
BERT-multilingual	168M	Geral	Multilíngue
XLM-RoBERTa-base	279M	Geral	Multilíngue
BERTimbau-large	335M	Geral	Português
XLM-RoBERTa-large	561M	Geral	Multilíngue
ALBERTINA-PTBR	900M	Geral	Português

Fonte: De autoria própria.

BERDOU (CAÇÃO *et al.*, 2022), tendo este sido pré-treinado utilizando textos do DOU. Esses modelos foram baseados no BERTimbau-base, uma versão simplificada do BERTimbau, contendo menos parâmetros do que o BERTimbau-large. Na Tabela 4.4 são resumidas as informações de cada LLM de arquitetura *encoder*, incluindo seu tamanho e domínio de pré-treinamento.

4.3.3 Modelos de Linguagem Grandes (*Large Language Models* - LLMs) baseados em arquitetura *decoder*

De forma a realizar uma análise comparativa entre LLMs *encoders* e *decoders*, foi realizado um experimento utilizando um LLM baseado em arquitetura *decoder*, especificamente o modelo Llama 2 (TOUVRON *et al.*, 2023b). Optou-se pela versão de 7 bilhões de parâmetros, dadas as especificações de hardware disponíveis, uma vez que modelos maiores demandam ainda mais recursos computacionais, indisponíveis nesta pesquisa. Foram realizados, inicialmente, experimentos utilizando *zero-shot*, *one-shot* e *few-shot learning*, de forma a utilizar a capacidade do modelo de inferir a classificação a partir da sua janela de contexto e pré-treinamento. Porém, os experimentos não obtiveram resultados satisfatórios.

Para garantir uma comparação mais precisa dos resultados, foi realizado o treinamento (*fine-tuning*) do modelo especificamente para o problema de classificação binária. Mesmo

com a utilização da menor versão modelo Llama 2 disponível, ainda assim foram utilizadas técnicas de ajuste eficiente de parâmetros (PEFT, do inglês, *Parameter-Efficient Fine-Tuning*) (MANGRULKAR *et al.*, 2022) que possibilitaram o treinamento do LLM de arquitetura *decoder* em uma única GPU. Para o experimento em questão, foi utilizado o método QLoRA (DETTMERS *et al.*, 2023). Essas técnicas mantêm os parâmetros obtidos a partir do pré-treinamento congelados, adicionando parâmetros extras que serão ajustados no processo de treinamento (*fine-tuning*). Dessa forma, o modelo mantém os pesos originais, reduzindo a carga computacional para a adaptação dos LLMs para cenários específicos, uma vez que não é necessário realizar o ajuste de todos os pesos do modelo.

Além disso, o modelo também passa por um processo de quantização (do inglês, *quantization*), reduzindo a precisão dos pesos do modelo de pontos flutuantes de 32 bits para, por exemplo, inteiros de 8 bits. Dessa forma, o modelo passa a ser mais eficiente, consumindo menos recursos, memória e com processamento mais rápido. Diante disso, é essencial encontrar um balanço na precisão utilizada, de forma a não afetar os resultados finais por conta de uma redução demasiada, o que pode acabar reduzindo a capacidade de representação do modelo.

4.4 Avaliação

Para analisar a eficácia dos modelos, foi levado em consideração o alto desbalanceamento inerente ao problema alvo, que envolve a classificação de texto tributário em publicações do DOU. As amostras negativas constituíram a classe majoritária, portanto, qualquer métrica calculada considerando a classe negativa foi severamente impactada. Suponha-se um classificador enviesado, no qual o resultado seja de que todos os registros são não tributários, ou seja, probabilidade 0 de ser tributário. Considerando o conjunto de testes descrito na Seção 4.2, onde dos 344.569 registros existentes, 344.254 não são tributários, as métricas como precisão e revocação para a classe negativa teriam valores extremamente altos, ou seja, próximos a 1. Por consequência, assim também ocorreria com o F1-Score para a classe negativa, dado o valor extremamente alto para a precisão e revocação, métricas que compõem o F1-Score.

Este impacto também é observado em métricas que levam ambas as classes, positiva

e negativa, em consideração. Dado que métricas variantes do F1-Score, como *Weighted-Average*, *Macro-Average* e *Micro-Average*, que possuem seu valor afetado de acordo com a proporção observada no desbalanceamento dos dados, também seriam enviesadas por conta do alto volume da classe negativa. A métrica de acurácia também representaria um valor extremamente alto, mascarando a incapacidade do modelo de exemplo de reconhecer sequer um único registro da classe positiva.

Como resultado desta análise, em caso da classe negativa possuir um grande número de acertos, o que se confirmou sendo o caso nos experimentos realizados neste trabalho, a classe positiva deveria ser o foco da análise e avaliação dos resultados obtidos de cada modelo treinado. Sendo assim, as seguintes métricas mostraram-se adequadas neste cenário de classificação binária altamente desbalanceado: a revocação e a precisão, ambas da classe positiva, e a Área Sob a Curva de Precisão-Revocação (PR-AUC, do inglês, *Precision-Recall Area Under the Curve*). Dessa forma, será possível analisar cada modelo e comparar os resultados de forma mais justa e assertiva.

É importante ressaltar que, apesar deste trabalho de pesquisa buscar a identificação da maior quantidade possível de atos tributários em meio a todas as publicações do Diário Oficial da União, o contexto é extremamente dificultoso e erros são naturais em aplicações de AM no mundo real. De forma a reduzir a quantidade de publicações a serem analisadas manualmente, além de ter um panorama da quantidade de publicações tributárias publicadas no mesmo dia da publicação, a métrica que se busca maximizar é a revocação, englobando, assim, a maior quantidade de publicações tributárias existentes. É evidente que deve haver um balanço nessa maximização, pois uma revocação de 1 acompanhado de uma precisão de 0 não resulta em ganho algum.

Ainda que os resultados obtidos a partir dos modelos treinados possam apresentar uma quantidade significativa de Falsos Positivos, ou seja, publicações classificadas como tributárias, mas não sendo, não configura necessariamente num resultado ruim. Desde que os Falsos Negativos, ou seja, publicações tributárias que deixaram de ser identificadas pelos modelos, sejam minimizados, é tolerável a existência de Falsos Positivos, desde que haja um balanço nessa quantidade. Ajustes no limiar de classificação entre a classe positiva e negativa podem ser um fator determinante na obtenção de um resultado relevante. Diante disso, além das métricas computadas para os valores de limiar de classificação padrão de 0,5, fo-

ram calculados os limiares de classificação que maximizem o F1-Score da classe positiva para cada um dos modelos treinados, buscando o melhor balanço possível de cada modelo entre precisão e revocação. O valor de limiar foi calculado utilizando a implementação da função *precision_recall_curve*[2] da biblioteca sklearn.

Além disso, de forma a oferecer uma solução adaptável a diferentes cenários, o objetivo do fluxo de processamento é de associar cada registro processado a uma probabilidade de ser de domínio tributário. Dessa forma, torna-se possível realizar uma análise dos registros classificados utilizando diferentes limiares de classificação, tendo como foco três principais limiares:

1. Publicações com uma alta probabilidade de possuir conteúdo tributário;
2. Publicações compreendidas num intervalo de média probabilidade de possuir conteúdo tributário; e
3. Publicações com uma baixa probabilidade de possuir conteúdo tributário.

Dessa maneira, especialistas na análise tributária podem realizar análises do conteúdo classificado tomando como base esses limiares de classificação. Assim, publicações cuja probabilidade de possuir conteúdo tributário esteja num intervalo de média probabilidade poderiam ser indicadas como passíveis de uma análise humana. Apesar do foco deste trabalho ser o de identificar o máximo de publicações tributárias possível de forma automática, o contexto extremamente desbalanceado pode impactar significativamente no resultado, com muitos falsos positivos ou negativos quando aplicado a um cenário real. Trabalhar então com diferentes limiares ou intervalos de classificação pode contribuir significativamente de forma a melhorar o desempenho dos modelos na fase de *Deployment* da metodologia CRISP.

4.5 Considerações Finais

Neste capítulo foi apresentada a metodologia utilizada nesta pesquisa contemplando os materiais utilizados, que representam o *corpus* de atos no domínio de direito tributário do DOU e os métodos, que são os modelos de AM escolhidos, bem como as métricas adotadas para

[2]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html

aferir eficiência e eficácia de tais modelos. No próximo capítulo, serão discutidos os experimentos realizados e uma avaliação de eficiência e eficácia dos modelos.

Capítulo 5

Avaliação Experimental

Neste capítulo, são apresentados os resultados obtidos dos experimentos realizados, bem como uma avaliação acerca do desempenho dos modelos de aprendizagem utilizados na classificação de atos tributários publicados no DOU.

O presente capítulo é dividido da seguinte forma: na Seção 5.1 é descrito o ambiente de execução dos experimentos; na Seção 5.2 são detalhados os hiperparâmetros selecionados para os modelos; na Seção 5.3 estão listados os resultados alcançados com os experimentos realizados, bem como uma discussão detalhada acerca destes resultados; por fim, na Seção 5.4, são contempladas as considerações finais acerca da etapa de avaliação experimental desta pesquisa.

5.1 Ambiente de Execução

Os experimentos conduzidos nesta pesquisa foram realizados num ambiente computacional controlado, objetivando prover confiabilidade aos resultados. Nas subseções a seguir serão descritas as configurações do ambiente utilizado.

5.1.1 Configurações de Hardware

Todos os experimentos foram realizados numa máquina servidora dedicada contendo a seguinte configuração de hardware:

- **CPU:** Intel Core i7 10700KF, 3.80 GHz, 8 núcleos e 16 *threads*;

- **RAM:** 64 GB DDR4 3200MHz;
- **GPU:** NVIDIA GeForce RTX 4090 24GB;
- **Armazenamento:** 2 TB NVMe SSD; e
- **Velocidade de Internet:** 100 Mbps.

5.1.2 Configurações de Software

A máquina servidora estava equipada com as seguintes configurações de software:

- **Sistema Operacional:** Pop!_OS 20.04 LTS;
- **Drivers CUDA:** 12.2;
- **Python:** 3.9.16; e
- **Principais Bibliotecas:**
  - Jupyter Lab 4.0.2;
  - TensorFlow 2.12.0;
  - PyTorch 2.0.1.

5.2 Seleção de Hiperparâmetros

De forma a selecionar os melhores hiperparâmetros para cada técnica aplicada à tarefa de classificação, cada solução adotada passou por uma seleção de hiperparâmetros otimizados ao problema. Para os modelos *baseline* SVM e XGBoost, foi utilizado o *framework* Optuna[1], que busca o valor que otimiza os resultados a partir de uma função objetivo. O *Tree-structured Parzen Estimator* (TPE) foi utilizado como heurística de amostragem dos hiperparâmetros utilizados em cada execução (AKIBA *et al.*, 2019). Como critério de parada das execuções (*pruning*), foi adotada a estratégia da mediana, interrompendo as execuções cujos resultados estivessem abaixo da mediana das execuções anteriores para uma mesma

[1] https://optuna.org/

Tabela 5.1: Resumo dos Hiperparâmetros utilizados na seleção do modelo SVM com Optuna e Hiperparâmetros selecionados, tanto para o conjunto de treinamento balanceado como para o desbalanceado.

Hiperparâmetro	**Espaço de Busca**	**Valor Selec. (Balan.)**	**Valor Selec. (Desba.)**
C	[min=1e-10, max=1e10, log=True]	1,250e-2	1,110e6
kernel	["linear "poly "rfb"]	"rfb"	"poly"
degree	[1, 2, 3]	3	3
gamma	["scale", "auto"]	"scale"	"scale"

Fonte: De autoria própria.

etapa (*step*). Os valores dos hiperparâmetros considerados na busca foram definidos previamente, estando listados na Tabela 5.1 os parâmetros respectivos ao SVM e na Tabela 5.2 os do XGBoost, juntamente dos valores selecionados para os experimentos utilizando ambos os conjuntos de dados.

Considerando-se a estratégia de AutoML empregada, o modelo resultante da seleção otimizada foi o *Passive Aggressive*, sendo um algoritmo de aprendizagem incremental, comumente utilizado em tarefas de aprendizagem de larga escala (KUMAR, 2023). Conforme sua própria nomenclatura, o algoritmo comporta-se passivamente frente a classificações corretas, não realizando alterações nos parâmetros, e agressivamente frente as incorretas, realizando alterações (CRAMMER *et al.*, 2006; KUMAR, 2023). A própria ferramenta de AutoML realizou a otimização de hiperparâmetros de forma automática, em conjunto com a seleção dos algoritmos utilizados, utilizando-se do método de otimização Bayesiana (FEURER *et al.*, 2015). Os modelos considerados na seleção e otimização pela ferramenta Auto-Sklearn estão listados na Tabela 5.3. Os hiperparâmetros otimizados selecionados para o modelo *Passive Aggressive* estão listados na Tabela 5.4.

Para os LLMs de arquitetura *encoder*, todos os hiperparâmetros utilizados nos experimentos foram determinados por meio de múltiplos experimentos realizados por modelo, conforme listado na Tabela 5.5. Foram empregadas dez épocas em todos os experimentos, tendo sido observado como um valor que levou à divergência no erro de validação em todos os modelos, extraindo assim o melhor resultado para cada combinação de hiperparâmetro possível. Quanto ao otimizador aplicado aos modelos, foi utilizado o AdamW, um dos mais utilizados na literatura (LOSHCHILOV; HUTTER, 2017; WAN *et al.*, 2023). O estado

Tabela 5.2: Resumo dos Hiperparâmetros utilizados na seleção do modelo XGBoost com Optuna e Hiperparâmetros selecionados, tanto para o conjunto de treinamento balanceado como para o desbalanceado.

Hiperparâmetro	**Espaço de Busca**	**Valor Selec. (Balan.)**	**Valor Selec. (Desba.)**
booster	["gbtree "gblinear "dart"]	"gblinear"	"gblinear"
lambda	[min=1e-8, max=1, log=true]	6,509e-5	5,353e-3
alpha	[min=1e-8, max=1, log=true]	2,584e-8	7,908e-6
updater	"shotgun"	"shotgun"	"shotgun"
top_k	"cyclic"	"cyclic"	"cyclic"
feature_selector	0	0	0
Para valor de booster igual a "gbtree" ou "dart":			
max_depth	[1, 2, 3, 4, 5, 6, 7, 8, 9]	-	-
eta	[min=1e-8, max=1, log=true]	-	-
gamma	[min=1e-8, max=1, log=true]	-	-
grow_policy	["depthwise "lossguide"]	-	-
Para valor de booster igual a "dart":			
sampler_type	[üniform "weighted"]	-	-
normalize_type	["tree "forest"]	-	-
rate_drop	[min=1e-8, max=1, log=true]	-	-
skip_drop	[min=1e-8, max=1, log=true]	-	-

Fonte: De autoria própria.

Tabela 5.3: Algoritmos de Classificação Considerados na Seleção do Auto-Sklearn.

Algoritmo de Classificação
AdaBoost (AB)
Bernoulli Naïve Bayes
Decision Tree (DT)
Extreml. Rand. Trees
Gaussian Naïve Bayes
Gradient Boosting (GB)
kNN
LDA
Linear SVM
Kernel SVM
Multinomial Naïve Bayes
Passive Aggressive
QDA
Random Forest (RF)
Linear Class. (SGD)

Fonte: (FEURER *et al.*, 2015)

Tabela 5.4: Resumo dos Hiperparâmetros selecionados para o classificador *Passive Aggressive*, tanto para o conjunto de treinamento balanceado como para o desbalanceado.

Hiperparâmetro	**Valor Selec. (Balan.)**	**Valor Selec. (Desba.)**
C	0,106	4,184
average	True	True
loss	squared_hinge	squared_hinge
max_iter	32	16
tol	4,414e-5	0,200e-1
warm_start	True	True

Fonte: De autoria própria.

Tabela 5.5: Hiperparâmetros utilizados na seleção dos modelos LLMs *encoder* e do modelo LLM *decoder*, o Llama 2.

Hiperparâmetro	**Espaço de Busca (LLMs *encoder*)**	**Espaço de Busca (LLM *decoder*)**
learning_rate	[5e-5, 3e-5, 1e-5, 5e-4, 3e-4, 1e-4, 5e-3, 3e-3, 1e-3]	[5e-5, 2e-4, 3e-4, 5e-4]
per_device_train_batch_size	[4, 8, 16]	[8, 16]
gradient_accumulation_steps	1	[4, 8]
per_device_eval_batch_size	[4, 8, 16]	[8, 16]

Fonte: De autoria própria.

de treinamento dos modelos (*checkpoints*) foram salvos para cada época em que houve aprimoramento durante a etapa de avaliação do modelo. Os modelos e hiperparâmetros que resultaram nas melhores métricas foram salvos, estando listados nas Tabelas 5.6 e 5.7, para o conjunto de dados balanceado e desbalanceado, respectivamente.

Tratando-se do LLM de arquitetura *decoder* utilizado, também foram realizados múltiplos experimentos, considerando as combinações dos hiperparâmetros também listados na Tabela 5.5. Os demais hiperparâmetros foram mantidos fixos, dado o alto custo computacional dos experimentos envolvendo o LLM Llama 2. É importante ressaltar que o processo de treinamento deste modelo foi realizado utilizando-se da aplicação de técnicas de quantização e PEFT, mais especificamente o QLoRA. Assim como nos LLMs *encoder*, para cada época de treinamento foi realizada a avaliação e, em caso de aprimoramento, armazenou-se o estado do modelo.

Por fim, para o Llama 2, foram utilizados no processo de *fine-tuning* os hiperparâmetros

Tabela 5.6: Resumo dos hiperparâmetros selecionados para os LLMs de arquitetura *encoder* utilizados nos experimentos no conjunto de dados balanceado.

Modelo	Épocas	*Batch Size*	*Batch Status*	*Learning Rate*	Otimizador	Tam. Máx. Seq.
BERTimbau-base	3	16	32	3e-5	AdamW	512 tokens
Legal-BERT	3	8	16	3e-5	AdamW	512 tokens
BERTikal	5	16	32	2e-5	AdamW	512 tokens
BERDOU	2	8	32	3e-5	AdamW	512 tokens
BERT-multilingual	4	8	32	3e-5	AdamW	512 tokens
XLM-RoBERTa-base	3	8	32	1e-5	AdamW	512 tokens
BERTimbau-large	3	16	32	3e-5	AdamW	512 tokens
XLM-RoBERTa-large	2	8	32	1e-5	AdamW	512 tokens
ALBERTINA-PTBR	4	4	16	1e-5	AdamW	512 tokens

Fonte: De autoria própria.

Tabela 5.7: Resumo dos hiperparâmetros selecionados para os LLMs de arquitetura *encoder* utilizados nos experimentos no conjunto de dados desbalanceado.

Modelo	Épocas	*Batch Size*	*Batch Status*	*Learning Rate*	Otimizador	Tam. Máx. Seq.
BERTimbau-base	3	16	32	3e-5	AdamW	512 tokens
Legal-BERT	5	8	32	1e-5	AdamW	512 tokens
BERTikal	10	16	32	1e-5	AdamW	512 tokens
BERDOU	4	16	32	3e-5	AdamW	512 tokens
BERT-multilingual	2	16	32	3e-5	AdamW	512 tokens
XLM-RoBERTa-base	3	16	32	1e-5	AdamW	512 tokens
BERTimbau-large	3	16	32	3e-5	AdamW	512 tokens
XLM-RoBERTa-large	4	16	32	1e-5	AdamW	512 tokens
ALBERTINA-PTBR	2	4	32	1e-5	AdamW	512 tokens

Fonte: De autoria própria.

Tabela 5.8: Hiperparâmetros do experimento utilizando Llama 2: treinamento, QLoRA e inferência (tanto para o conjunto balanceado como para o desbalanceado).

Hiperparâmetro	**Valor**
Batch Size	16
Otimizador	paged_adamw_32bit
Learning Rate	2e-5
Máx. Tam. Seq.	512 Tokens
Épocas	10
Quantização - Método	8bit + FP8
Quantização - *compute dtype*	bf16
Quantização - *nested quantization*	True
QLoRA - *Alpha*	16
QLoRA - *Dropout*	0,1
QLoRA - r	64
QLoRA - *target_modules*	q_proj; v_proj
Inferência - *temperature*	0,8
Inferência - top_p	0,5
Inferência - max_new_tokens	5

Fonte: De autoria própria.

elencados na Tabela 5.8. O *fine-tuning* foi realizado com o auxílio do QLoRA (DETTMERS *et al.*, 2023). Para o experimento no formato de *few-shot learning*, sem realização de ajuste de parâmetros e pesos, os mesmos valores dos hiperparâmetros de inferência listados na Tabela 5.8 foram utilizados.

5.3 Resultados e Discussão

Frente a todos os experimentos realizados, o modelo BERTimbau-large utilizado demonstrou o desempenho de maior destaque, superando os demais modelos em termos de PR-AUC, tanto no cenário balanceado como no desbalanceado, conforme resultados listados na Tabela 5.9. Os modelos tradicionais, especificamente SVM, XGBoost e *Passive Aggressive*, exibiram resultados inferiores aos LLMs de arquitetura *encoder*. Ademais, esses modelos apresentaram valores em torno de 0,5 para a métrica PR-AUC, caracterizando uma performance subótima, exceto para o XGBoost, no cenário balanceado, e o *Passive Aggressive*, no

Tabela 5.9: Resultados dos experimentos em termos de PR-AUC, tanto para treinamento balanceado como desbalanceado (LLMs ordenados em ordem crescente pela quantidade de parâmetros; Diferença $= [(ValorFinal - ValorInicial)/ValorInicial] * 100$).

Modelo	PR-AUC (Balanceado)	PR-AUC (Desbalanceado)	Diferença (%)
SVM	0,403822	0,512648	**26,95**
XGBoost	0,619059	0,407853	-34,12
Passive Aggressive (AutoML)	0,535633	0,643388	20,12
BERT-multilingual	0,544775	0,681180	25,04
BERTimbau-base	0,616078	0,726627	17,94
Legal-BERT	0,646020	0,688137	6,52
BERTikal	0,666752	0,715527	7,32
BERDOU	0,673919	0,725577	7,67
XLM-RoBERTa-base	0,624070	0,732478	17,37
BERTimbau-large	**0,719511**	**0,849973**	18,13
XLM-RoBERTa-large	0,654323	0,737347	12,69
ALBERTINA	0,666296	0,761102	14,23

Fonte: De autoria própria.

cenário desbalanceado. O modelo BERTimbau-large ajustado no conjunto de treinamento desbalanceado superou as abordagens que utilizaram o SVM (balanceado) e o XGBoost (desbalanceado) com uma diferença de quase o dobro em termos de PR-AUC, mesmo após a otimização dos hiperparâmetros. Ao contrastar a métrica PR-AUC entre todos os LLMs empregados na pesquisa, o modelo BERTimbau-large desbalanceado destacou-se com uma margem de, pelo menos, 11,67% superior aos valores obtidos pelos demais modelos. Os resultados em termos de PR-AUC, tanto para o cenário balanceado quanto para o desbalanceado, bem como a diferença de um cenário para o outro, estão listados na Tabela 5.9.

As demais métricas: precisão, revocação e F1-Score, também foram computadas, todas para a classe positiva, juntamente com o limiar de classificação otimizado para o F1-Score da classe positiva no conjunto de validação. Esses resultados encontram-se listados na Tabela 5.10, para o cenário balanceado, e na Tabela 5.11, para o desbalanceado. Para os experimentos realizados utilizando o modelo Llama 2, representante dos LLMs *decoder* nesta pesquisa, os resultados encontram-se listados na Tabela 5.12 nas três abordagens utilizadas: *few-shot*, treinamento balanceado e treinamento desbalanceado. Para o modelo Llama 2 não foram

computadas as métricas de PR-AUC e também não foi calculado o limiar de classificação otimizado, dado que a saída do modelo é textual, diferentemente dos demais modelos cuja saída representa uma probabilidade.

Tabela 5.10: Resultados dos experimentos utilizando técnicas tradicionais de PLN e LLMs de arquitetura *encoder* para o conjunto de dados balanceado.

Modelo	PR-AUC	Limiar de Classificação	Precisão (+)	Revocação (+)	F1-Score (+)
SVM	0,403822	0,5	0,06	0,75	0,11
		0,564391	0,72	0,28	0,41
XGBoost	0,619059	0,5	0,04	0,97	0,08
		0,973072	0,64	0,50	0,56
Passive Aggressive (AutoML)	0,535633	0,5	0,05	**0,99**	0,10
		0,745099	0,57	0,38	0,46
BERT-multilingual	0,544775	0,5	0,19	0,78	0,31
		0,979090	0,53	0,49	0,51
BERTimbau-base	0,616078	0,5	0,11	0,93	0,20
		0,996901	0,55	0,57	0,56
Legal-BERT	0,646020	0,5	0,12	0,90	0,21
		0,997131	0,50	0,70	0,58
BERTikal	0,666752	0,5	0,14	0,91	0,25
		0,998716	0,58	0,67	0,62
BERDOU	0,673919	0,5	0,14	0,97	0,24
		0,991149	**0,81**	0,54	**0,65**
XLM-RoBERTa-base	0,624070	0,5	0,10	0,95	0,19
		0,996036	0,40	0,78	0,52
BERTimbau-large	**0,719511**	0,5	0,15	**0,98**	0,25
		0,991409	0,49	0,82	0,61
XLM-RoBERTa-large	0,654323	0,5	0,17	0,95	0,29
		0,942997	0,69	0,61	0,64
ALBERTINA	0,666296	0,5	0,19	0,90	0,31
		0,993000	**0,81**	0,54	**0,65**

Fonte: De autoria própria.

Tabela 5.11: Resultados dos experimentos utilizando técnicas tradicionais de PLN e LLMs de arquitetura *encoder* para o conjunto de dados desbalanceado.

Modelo	PR-AUC	Limiar de Classificação	Precisão (+)	Revocação (+)	F1-Score (+)
SVM	0,512648	0,5	0,15	0,79	0,26
		0,999999	0,70	0,38	0,49
XGBoost	0,407853	0,5	0,42	0,43	0,42
		0,526159	0,55	0,37	0,44
Passive Aggressive (AutoML)	0,643388	0,5	0,10	0,95	0,18
		0,741666	0,59	0,56	0,57
BERT-multilingual	0,681180	0,5	0,28	0,81	0,42
		0,979596	0,72	0,62	0,66
BERTimbau-base	0,726627	0,5	0,12	**0,97**	0,21
		0,998372	0,79	0,59	0,67
Legal-BERT	0,688137	0,5	0,29	0,78	0,42
		0,998838	0,92	0,50	0,64
BERTikal	0,715527	0,5	0,25	0,85	0,39
		0,991495	0,90	0,51	0,65
BERDOU	0,725577	0,5	0,13	**0,97**	0,23
		0,999415	0,72	0,69	0,70
XLM-RoBERTa-base	0,732478	0,5	0,31	0,86	0,46
		0,994754	0,85	0,58	0,69
BERTimbau-large	**0,849973**	0,5	0,64	**0,97**	**0,77**
		0,825867	0,70	0,86	**0,77**
XLM-RoBERTa-large	0,737347	0,5	0,51	0,78	0,62
		0,981493	**0,93**	0,56	0,70
ALBERTINA	0,761102	0,5	0,16	0,92	0,28
		0,996593	0,87	0,62	0,72

Fonte: De autoria própria.

Tabela 5.12: Resultados dos experimentos utilizando modelo Llama 2.

Estratégia de Treinamento	Precisão (+)	Revocação (+)	F1-Score (+)
Few-shot (3 inst.)	0,00	1,00	0,00
Fine-tuning + QLoRA (Balan.)	0,05	0,92	0,09
Fine-tuning + QLoRA (Desba.)	0,08	0,87	0,15

Fonte: De autoria própria.

5.3.1 Comparação entre Resultados Obtidos a partir dos Conjuntos de Treinamento Balanceado e Desbalanceado

Nos resultados obtidos, considerando os modelos treinados a partir de ambos os conjuntos de treinamento considerados, o balanceado e o desbalanceado, observou-se uma melhoria nos resultados na maioria dos modelos desbalanceados, conforme as métricas listadas na Tabela 5.9. Para os cenários em que houve melhoria, o incremento médio nos resultados foi de 15,82% em termos de PR-AUC. O modelo que apresentou o maior incremento foi o SVM, com um valor de 26,95%. Apenas o modelo XGBoost não se beneficiou da inserção de desbalanceamento nos dados de treinamento, sendo observada uma redução de 34,12% em termos de PR-AUC. Dado que o referido modelo possui estratégias internas para o tratamento de conjuntos de dados desbalanceados, os resultados obtidos a partir da amostra de treinamento utilizada abrem margem para investigação futura. Seu resultado, em termos de PR-AUC, comparou-se ao BERTimbau-base balanceado, mostrando-se como uma opção interessante num cenário balanceado e com uma pequena quantidade de dados de treinamento.

A partir da comparação dos demais resultados entre ambas as estratégias de balanceamento, listados nas Tabelas 5.10 e 5.11, temos a ocorrência dos maiores valores de revocação nos experimentos com os conjuntos balanceados. No cenário balanceado, para o limiar de classificação de 0,5, grande parte dos modelos obtiveram altos valores de revocação, sendo estes iguais ou acima de 0,90 em 10 dos 12 modelos listados na Tabela 5.10. Entretanto, esses valores foram acompanhados de uma baixa precisão, tendo sido 0,19 o maior valor dentre os 10 modelos. Os modelos treinados a partir do conjunto balanceado, apesar de capturar bem os Verdadeiros Positivos no conjunto de testes, apresentaram um alto índice de Falsos Positivos, o que resultou na baixa precisão dos modelos.

Entretanto, ao analisar os resultados a partir dos experimentos no conjunto de dados desbalanceado, presentes na Tabela 5.11, observou-se uma redução na revocação e um aumento na precisão dos modelos em geral. Também verificou-se um aumento de F1-Score e PR-AUC, caracterizando assim um melhor resultado. Como consequência, pode-se concluir que os modelos treinados no conjunto de dados desbalanceado demonstraram uma menor tendência a apontar Falsos Positivos. Tomando como exemplo o BERTimbau-large, melhor modelo em ambos os cenários de balanceamento, esse comportamento torna-se ainda mais

evidente. Ao comparar o modelo balanceado com o desbalanceado, apesar da redução em 0,01 na revocação, a precisão saltou de 0,14 para 0,64 nos resultados para o modelo desbalanceado. Evidenciando assim, que a inserção de mais elementos negativos no conjunto de treinamento reduziu a quantidade de Falsos Positivos, que em termos quantitativos caiu de 1.822 para 169. O valor máximo da precisão em ambos os cenários também colabora para esta constatação, dado que no balanceado foi de 0,81, contra 0,93 do cenário desbalanceado.

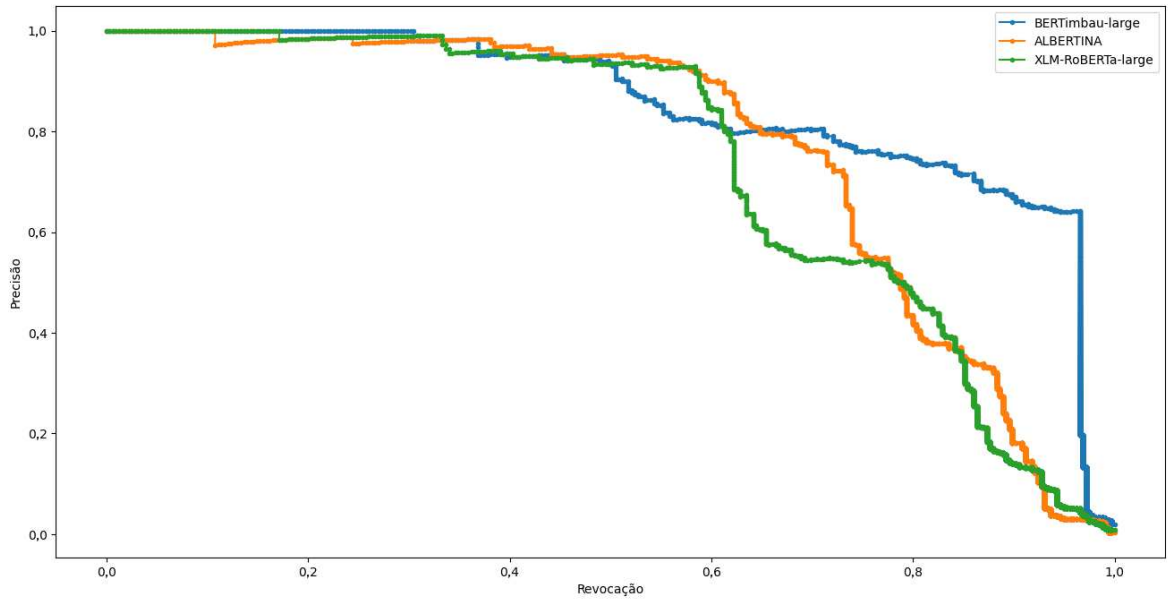
Sendo assim, para o cenário desta pesquisa, a inserção de um leve desbalanceamento no conjunto de treinamento dos modelos mostrou-se benéfica para a obtenção de melhores resultados, quando comparado a uma distribuição balanceada no conjunto de treinamento utilizando a estratégia de *undersampling*. Dado o desempenho superior, as análises realizadas nas subseções seguintes terão maior enfoque nos resultados obtidos a partir dos modelos treinados utilizando o conjunto desbalanceado, conforme as métricas detalhadas na Tabela 5.11.

5.3.2 Análise da Curva de Precisão-Revocação e Área Sob a Curva

Ao analisar as curvas de precisão-revocação dos três modelos mais eficientes provenientes dos experimentos, quais sejam: BERTimbau-large, ALBERTINA e XLM-RoBERTa-large, conforme representado na Figura 5.1, é possível verificar o comportamento desses modelos em resposta a diferentes limiares de classificação. No que concerne o modelo BERTimbau-large, nota-se que, para valores elevados de revocação, a precisão mantém-se em níveis elevados, ocorrendo leves declínios no gráfico à medida que os valores do limiar de classificação aumentam. Somente em valores de revocação próximos a 1,0 observa-se uma abrupta redução na curva, destacando a elevada qualidade do classificador obtido. Os limiares de classificação aumentam de maneira proporcional ao valor do revocação, da esquerda para a direita.

Ao examinar as curvas dos modelos ALBERTINA e XLM-RoBERTa-large, também representados na Figura 5.1, nota-se uma acentuada diminuição na precisão para valores de revocação superiores a 0,6. Conclui-se, portanto, que, para elevados valores de revocação, ambos os modelos não conseguiram manter uma precisão elevada entre os registros classificados, resultando em um número significativo de Falsos Positivos. As métricas associadas ao modelo ALBERTINA, utilizando o limiar de classificação de 0,5, corroboram com essa

Figura 5.1: Curvas de Precisão-Revocação dos três melhores modelos obtidos (em termos de PR-AUC): BERTimbau-large, ALBERTINA e XLM-RoBERTa-large.

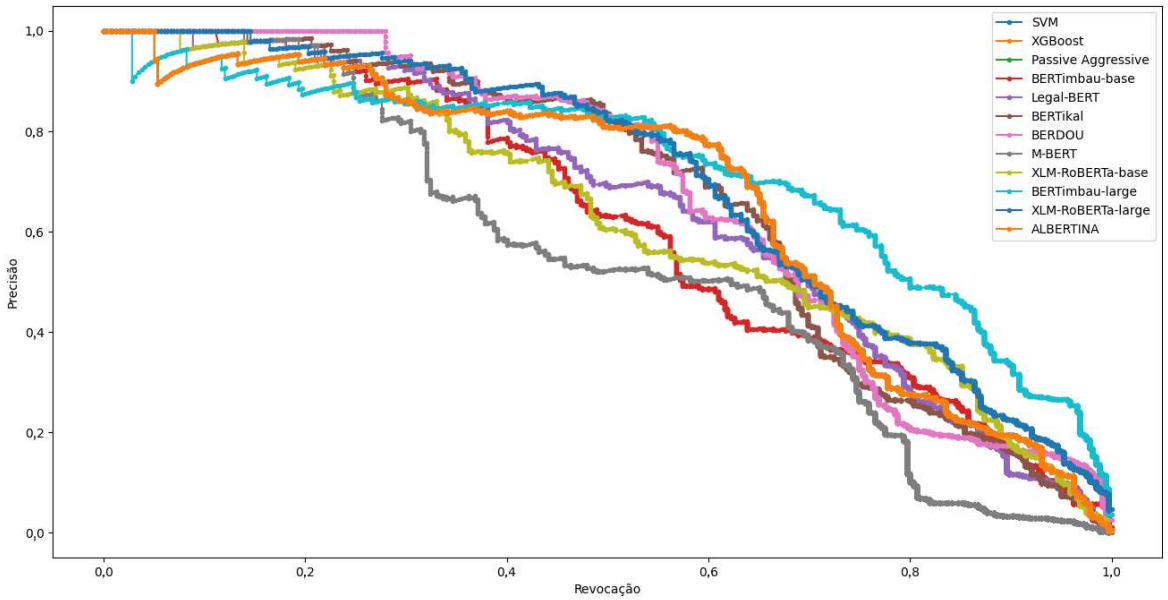


Fonte: De autoria própria.

conclusão, uma vez que, para um revocação de 0,92, a precisão alcançou 0,16. No caso do XLM-RoBERTa-large, apesar de ter exibido a maior precisão dentre todos os modelos analisados (0,93), tal precisão estava associada a um revocação de 0,56, aproximando-se do valor de 0,6 mencionado. Embora ambos os modelos tenham apresentado resultados superiores aos demais modelos testados, os mesmos obtiveram desempenhos inferiores quando comparados aos resultados obtidos com o BERTimbau-large.

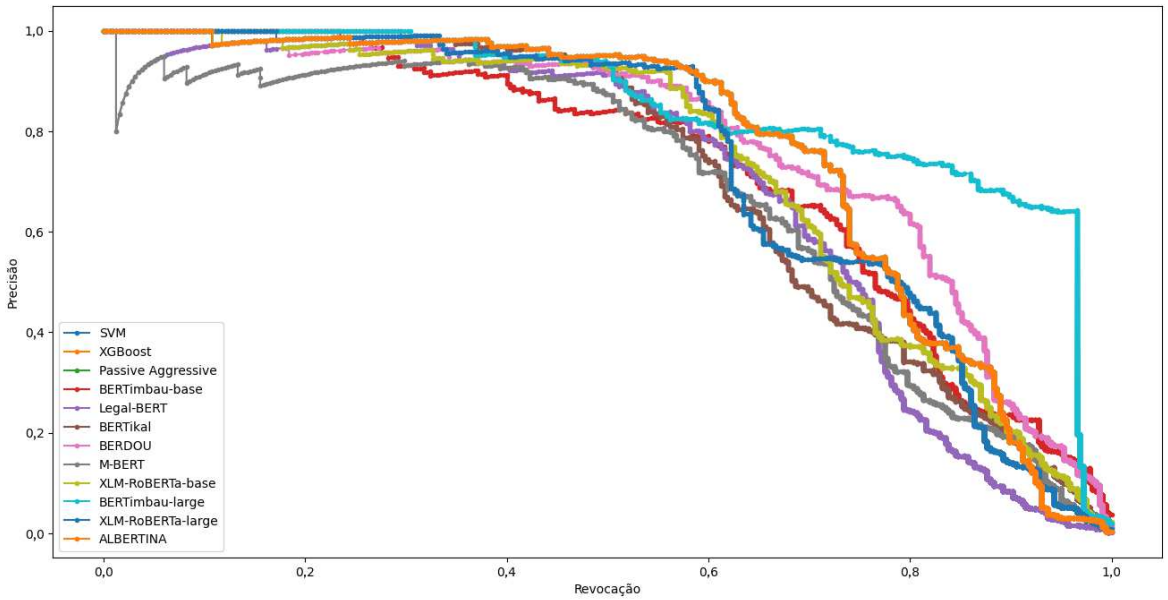
De forma a comparar os demais resultados, nas Figuras 5.2 e 5.3 estão dispostas as curvas de precisão-revocação de todos os modelos *baseline*, quais sejam: SVM, XGBoost e *Passive Aggressive*, e LLMs *encoder*. A Figura 5.2 contém os modelos balanceados, enquanto a Figura 5.3 os desbalanceados. O comportamento das curvas evidenciam o melhor desempenho do modelo BERTimbau-large desbalanceado frente aos resultados dos demais modelos utilizados nos experimentos.

Figura 5.2: Curvas de Precisão-Revocação de todos os modelos *baseline* e LLMs de arquitetura *encoder* (conjunto de treinamento balanceado).



Fonte: De autoria própria.

Figura 5.3: Curvas de Precisão-Revocação de todos os modelos *baseline* e LLMs de arquitetura *encoder* (conjunto de treinamento desbalanceado).



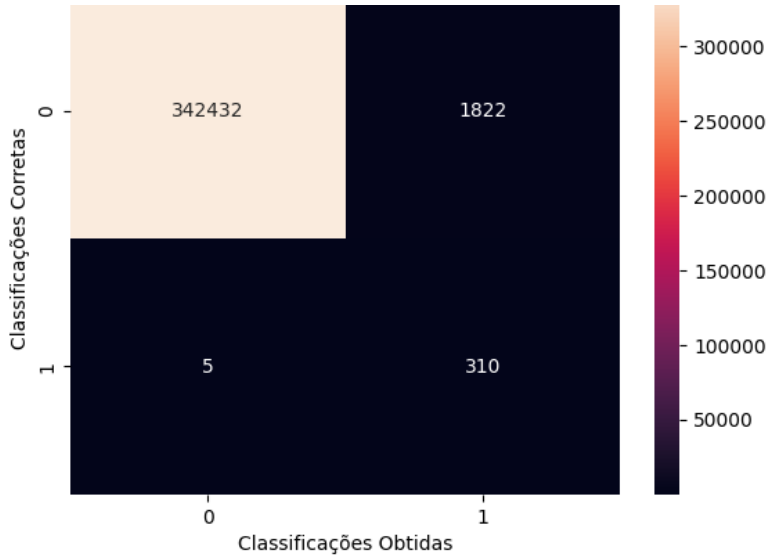
Fonte: De autoria própria.

5.3.3 Matriz de Confusão do BERTimbau-large e Limiares de Classificação

A matriz de confusão associada ao modelo BERTimbau-large ajustado balanceado está apresentada na Figura 5.4, tendo sido calculada mediante um limiar de classificação de 0,5, enquanto a Figura 5.5 apresenta a do modelo desbalanceado. O modelo BERTimbau-large desbalanceado demonstrou o melhor desempenho na classificação de publicações tributárias, registrando um PR-AUC de aproximadamente 0,85 e uma revocação de 0,97 para a classe positiva. Em termos quantitativos, isso se traduziu em 304 das 315 publicações tributárias sendo corretamente classificadas em um conjunto de dados que engloba um total de 344.569 publicações processadas.

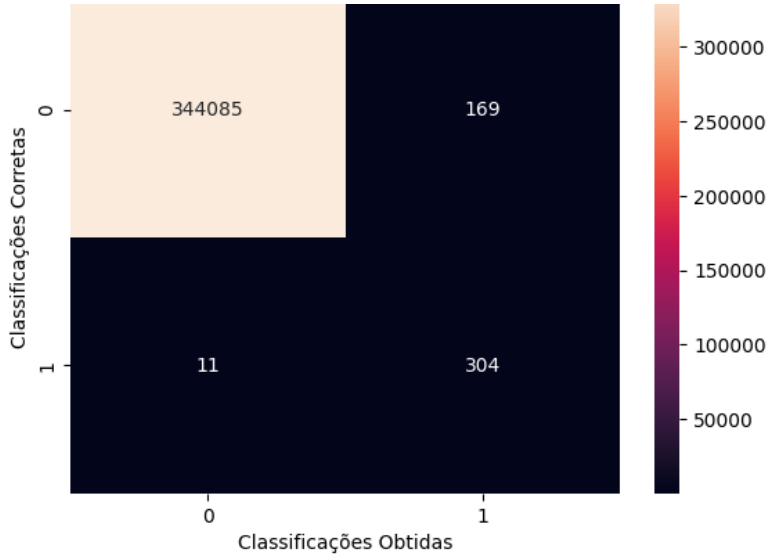
Apesar do modelo BERTimbau-large balanceado ter identificado mais registros positivos, alcançando uma revocação de 0,98, o que implicou em 310 das 315 publicações identificadas, 1.822 registros falsos positivos foram observados, contra apenas 169 do modelo desbalanceado. Sendo assim, observou-se uma maior precisão no modelo desbalanceado frente a uma leve redução em termos de revocação, representando em termos gerais um resultado superior.

Figura 5.4: Matriz de confusão do modelo BERTimbau-large balanceado (limiar de classificação de 0,5; 0 = não tributário; 1 = tributário).



Fonte: De autoria própria.

Figura 5.5: Matriz de confusão do modelo BERTimbau-large desbalanceado (limiar de classificação de 0,5; 0 = não tributário; 1 = tributário).



Fonte: De autoria própria.

A otimização dos limiares de classificação, realizada com base no F1-Score para a classe positiva, utilizando dados do conjunto de validação, teve como objetivo a minimização das classificações incorretas de cada classificador. Como resultado, uma considerável parcela dos limiares calculados apresentou valores próximos a 1,0, evidenciando a propensão dos modelos treinados, especialmente aqueles de arquitetura *transformer* baseada em *encoder*, a atribuir probabilidades elevadas na classificação de registros Verdadeiros Positivos. Ou seja, os modelos treinados demonstraram uma alta confiança na maioria das classificações positivas, conforme as distribuições das probabilidades atribuídas pelos modelos aos registros exibidas no Apêndice C. O BERTimbau-large ajustado, melhor modelo obtido, obteve um limiar ligeiramente inferior em comparação aos demais modelos de arquitetura semelhante, embora ainda elevado, sendo um valor aproximado de 0,83.

5.3.4 Comparação de Desempenho de Modelos Pré-treinados: Domínio Geral, Específico e Multilíngue

Uma análise pertinente, a ser conduzida com base nos resultados apresentados, envolve a comparação entre LLMs de domínio geral, e.g. BERTimbau-base, e aqueles especificamente voltados para domínios particulares, e.g. Legal-BERT, BERTikal e BERDOU. Ao

considerar modelos com tamanhos de parâmetros equivalentes para assegurar uma avaliação equilibrada, verificou-se que os modelos pré-treinados para domínios específicos exibiram desempenho semelhante ou inferior frente a modelos pré-treinados em domínio geral. A métrica PR-AUC, conforme apresentado na Tabela 5.11, evidencia essa relação entre os modelos. Considerando esse contexto, pode-se inferir que o pré-treinamento específico de LLMs não teve uma influência decisiva na melhoria dos resultados dentro do domínio específico abordado por esta pesquisa.

Uma outra análise realizada foi relacionada à utilização de modelos multilíngues. A análise dos modelos pré-treinados em português revelou um desempenho superior em comparação com todos os modelos multilíngues equivalentes, sendo estes o BERT-multilingual, XLM-RoBERTa-base e XLM-RoBERTa-large. Além disso, o modelo BERTimbau-large demonstrou resultados superiores quando contrastado com o modelo ALBERTINA, que possui quase três vezes mais parâmetros. Esse resultado questiona a crença convencional de que uma maior quantidade de parâmetros em LLMs inevitavelmente resulta em um desempenho aprimorado, destacando que tal suposição não é universalmente válida.

5.3.5 LLM de Arquitetura *Encoder* - Llama 2

Os experimentos conduzidos com o Llama 2, o maior dos modelos considerados nesta pesquisa, também contribuíram com evidências de que uma maior quantidade de parâmetros não implica necessariamente resultados superiores, conforme as métricas listadas na Tabela 5.12. No contexto do problema-alvo, o desempenho do modelo Llama 2 não foi satisfatório, sendo superado por todos os outros modelos empregados nos experimentos, os quais adotaram um limiar de classificação otimizado para o F1-Score da classe positiva. Além de apresentar resultados inferiores, o custo associado ao treinamento e processamento de todo o conjunto de testes foi consideravelmente elevado, a ser detalhado na seção subseção 5.3.6.

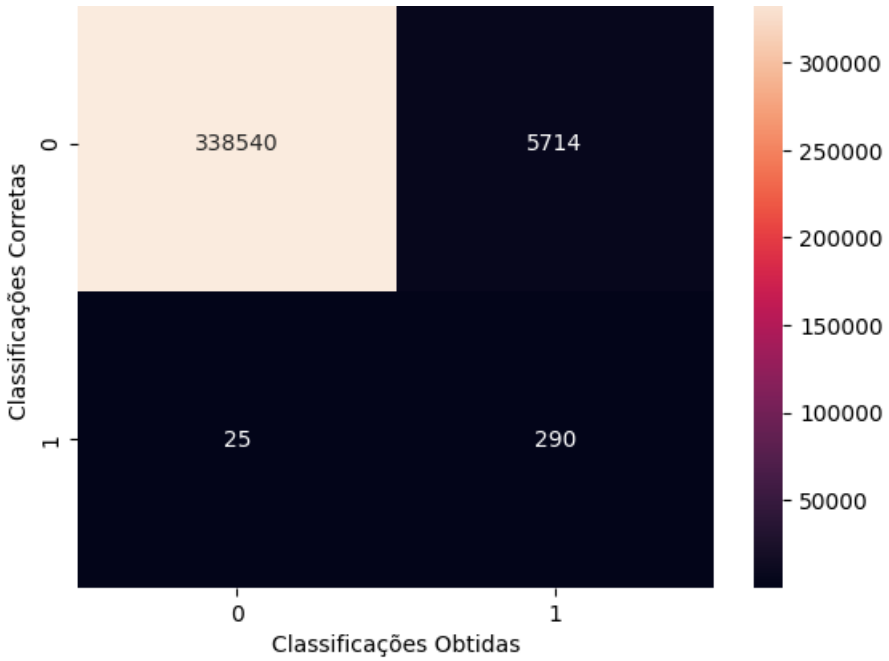
A aplicação do modelo Llama 2 ocorreu por meio de duas abordagens: inferência utilizando *few-shot learning* e *fine-tuning* com a utilização do QLoRA (DETTMERS *et al.*, 2023), esta última tendo sido aplicada em ambos os conjuntos de treinamento construídos. Foi empregada a otimização de quantização de parâmetros do modelo Llama 2 em todas as estratégias utilizadas. Utilizando-se da abordagem *few-shot*, estando esta associada a três exemplos, constatou-se que o Llama 2 não foi capaz de inferir os resultados de maneira satis-

fatória, conforme evidenciado pelos resultados na Tabela 5.12. Com menos de três exemplos na estratégia *few-shot*, as saídas esperadas "positivo" e "negativo", representando ambas as classes, não foram adequadamente mapeadas pelo modelo, resultando, por vezes, em textos completamente discrepantes do esperado. Mesmo com a realização de alterações no prompt utilizado na entrada, sendo este um fator de influência em abordagens *few-shot*, verificou-se o mesmo comportamento. Diante dessa constatação, não foi possível obter métricas conclusivas a partir de estratégias com menos de três exemplos. Não foram empregados mais de três exemplos nesse experimento dadas as limitações do hardware utilizado, ocasionando sobrecarga de memória e erros de execução nas tentativas realizadas.

No experimento de inferência utilizando a abordagem *few-shot*, todos os 344.569 registros do conjunto de testes foram processados, onde os exemplos *few-shot* consistiram em dois registros negativos e um positivo. Entretanto, o Llama 2 não foi capaz de capturar as características de um ato tributário de forma a classificá-lo corretamente, dado que o modelo atribuiu a classe positiva a todos os registros, isso considerando o conjunto de testes utilizado em conjunto com o *prompt* informado. Diante da ausência de resultados significativos dessa abordagem, conforme resultados apresentados na Tabela 5.12, não foram realizados experimentos adicionais utilizando-se de *few-shot learning*, dado seu alto custo de processamento.

Para o experimento utilizando o modelo Llama 2 ajustado no conjunto balanceado, observou-se uma melhoria nas métricas frente ao obtido com a estratégia *few-shot*, com precisão, revocação e F1-Score registrando valores de 0,05, 0,92 e 0,09, respectivamente. Houve aumento em termos de F1-Score quando comparado com o modelo treinado utilizando-se do conjunto de dados desbalanceado, que obteve o valor de 0,15. Observou-se também uma redução em termos de revocação para 0,87 e um aumento na precisão para 0,08, comportamento semelhante ao observado nos LLMs de arquitetura *encoder*. Contudo, os valores absolutos para cada uma das métricas ainda ficaram aquém quando comparados aos observados na maioria dos LLMs *encoder*. Dessa forma, nos experimentos conduzidos com o Llama 2, o modelo não apresentou resultados competitivos frente aos demais modelos utilizados. Apesar da alta revocação, o modelo obteve uma precisão baixa, com uma alta ocorrência de Falsos Positivos. As métricas obtidas a partir dos experimentos estão dispostas na Tabela 5.12. Além disso, as Figuras 5.6 e 5.7 contêm os resultados na forma de uma matriz de confusão para os conjuntos balanceado e desbalanceado, respectivamente.

Figura 5.6: Matriz de confusão dos resultados do experimento balanceado com o modelo Llama 2 utilizando a estratégia de *fine-tuning* com modelo quantizado e QLoRA (0 = não tributário; 1 = tributário).



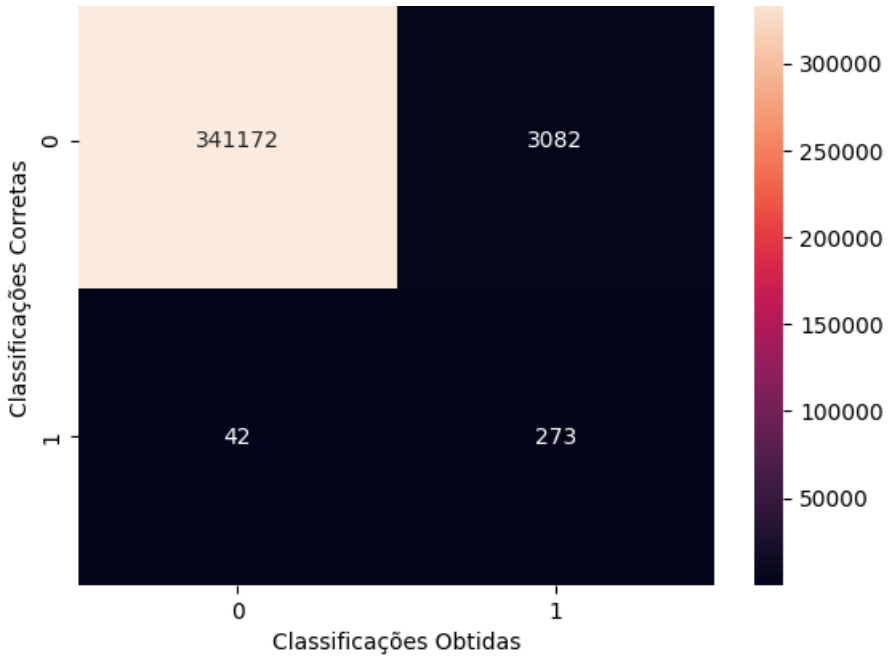
Fonte: De autoria própria.

Nos experimentos que incorporaram o modelo Llama 2, não foram computados os valores correspondentes à métrica PR-AUC, tampouco foi determinado o limiar de classificação otimizado. Em virtude da natureza generativa do LLM baseado em arquitetura *decoder*, a saída obtida é um texto gerado com base no *prompt* fornecido, ao invés de uma probabilidade de classificação, como é observado nos demais modelos. Consequentemente, para cada entrada, o modelo gerou um texto associado ao rótulo da classe atribuída, seja positiva ou negativa. Além disso, diante dos resultados inferiores do Llama 2 em comparação com os dos demais modelos, é importante salientar que o *prompt* utilizado, o *fine-tuning* com QLoRA e a quantização do modelo podem ter sido fatores de influência. Sendo assim, há margem para a realização de estudo mais aprofundado, inclusive utilizando uma maior quantidade de parâmetros, o que não se mostrou viável nesta pesquisa.

5.3.6 Análise de Eficiência dos Modelos

A análise da eficiência dos modelos foi aferida através do tempo de execução dos mesmos. Ao analisar os tempos de processamento para treinamento e avaliação utilizando o conjunto

Figura 5.7: Matriz de confusão dos resultados do experimento desbalanceado com o modelo Llama 2 utilizando a estratégia de *fine-tuning* com modelo quantizado e QLoRA (0 = não tributário; 1 = tributário).



Fonte: De autoria própria.

de testes em cada modelo, listados na Tabela 5.13, bem como nas Figuras 5.8 e 5.9, tornou-se evidente a disparidade entre os modelos Llama 2 e LLMs baseados em arquitetura *encoder* em termos de tempo de processamento. A métrica para o modelo Llama 2 na abordagem *few-shot* não foi adicionada ao gráfico da Figura 5.8, dada a ausência de métricas para treinamento e sua alta disparidade frente às demais métricas para o tempo de teste, que totalizou 56h, o que prejudicaria a visualização do gráfico. A Figura 5.9 é focada apenas nos tempos de teste, englobando também o tempo obtido com o Llama 2 *few-shot*.

Também foi possível observar o impacto da quantidade de parâmetros dos LLMs nos tempos de processamento, onde modelos maiores implicaram em maiores tempos de processamento, tanto para treinamento como para teste. A exemplo dos LLMs de arquitetura *encoder*, em que o tempo de processamento associado ao ALBERTINA foi mais de três vezes superior ao observado no BERTimbau-large, um modelo quase três vezes menor. Observou-se também que os tempos de execução para a etapa de testes nos LLMs foram próximos aos tempos de treinamento, etapa mais custosa do processo devido ao ajuste dos pesos dos modelos. A grande quantidade de registros presentes no conjunto de testes, compreendendo

Tabela 5.13: Tempo de processamento para treinamento e avaliação utilizando o conjunto de testes (o tempo computado para o *Passive Aggressive* também envolve a seleção de modelos e otimização de hiperparâmetros).

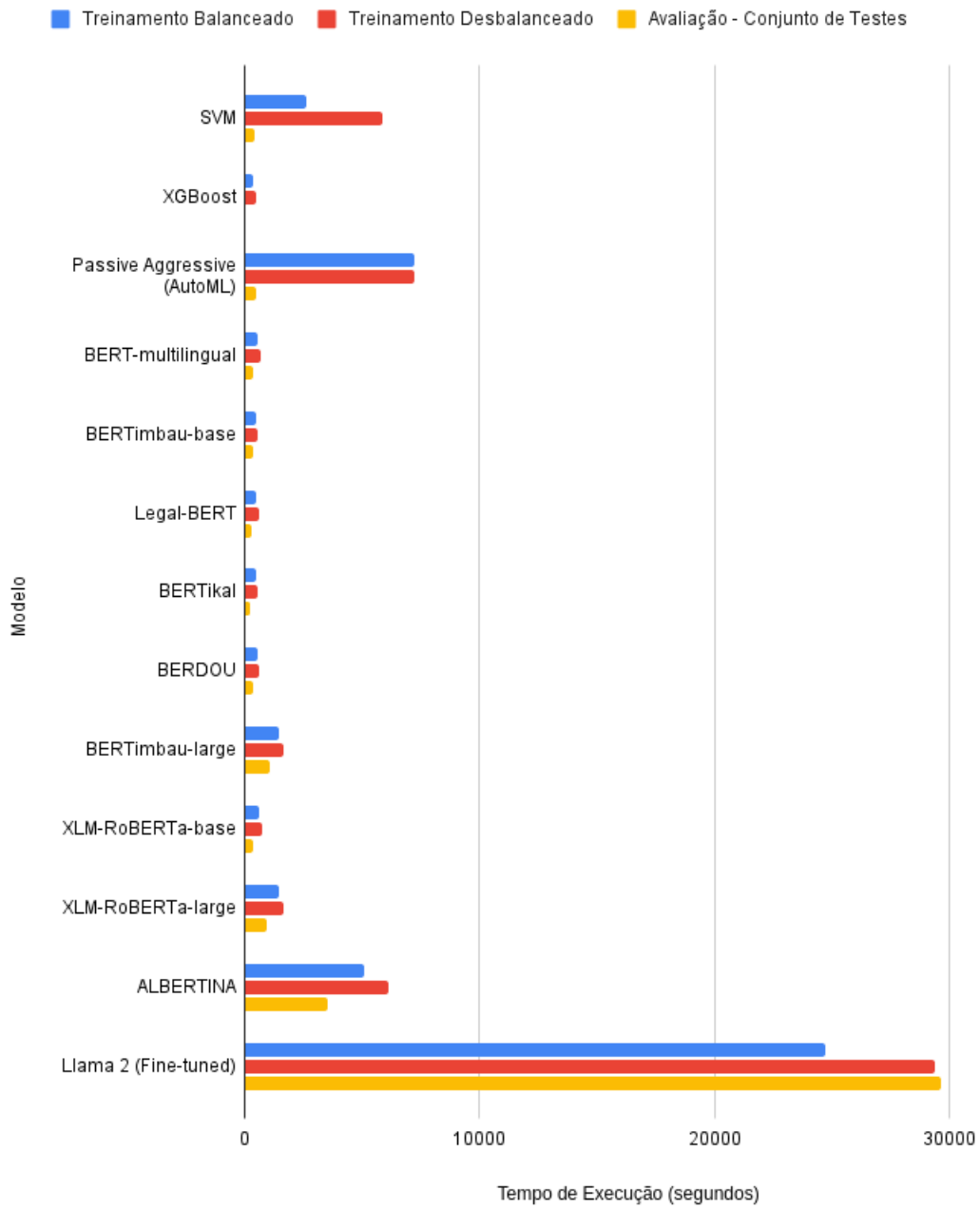
Modelo	Treinamento Balan. (segundos)	Treinamento Desba. (segundos)	Avaliação - Conj. de Testes (segundos)
SVM	2.631	5.843	401
XGBoost	357	479	0,04
Passive Aggressive (AutoML)	7.217	7.199	506
BERT-multilingual	566	659	351
BERTimbau-base	494	543	357
Legal-BERT	510	607	312
BERTikal	478	542	352
BERDOU	530	609	351
BERTimbau-large	1.440	1.658	1.052
XLM-RoBERTa-base	581	708	328
XLM-RoBERTa-large	1.451	1.670	927
ALBERTINA	5.094	6.100	3.539
Llama 2 (*Few-shot*)	-	-	202.560
Llama 2 (*Fine-tuned*)	24.694	29.375	29.648

Fonte: De autoria própria.

aproximadamente 350 mil registros, pode ter levado a essa situação.

No que diz respeito ao processo de treinamento do Llama 2, foram demandadas aproximadamente 7 horas para o modelo balanceado e mais de 8 horas para o desbalanceado, embora tenha sido empregado o método QLoRA para aprimorar a eficiência. Subsequentemente, o processo de inferência no conjunto de dados de teste completo usando Llama 2 ajustado com QLoRA consumiu 8 horas, representando 14% do tempo total utilizado pela abordagem *few-shot*. A abordagem utilizando o QLoRA, além de realizar o ajuste de parâmetros, dispensou a utilização de exemplos no contexto de inferência, reduzindo assim o custo de processamento, conforme observado ao comparar o tempo de execução entre as abordagens no Llama 2. Em contraste, considerando o BERTimbau-large desbalanceado como referência, que foi o melhor modelo obtido, o processo de *fine-tuning* foi concluído em cerca de 30 minutos, com 17 minutos adicionais para classificar todo o conjunto de da-

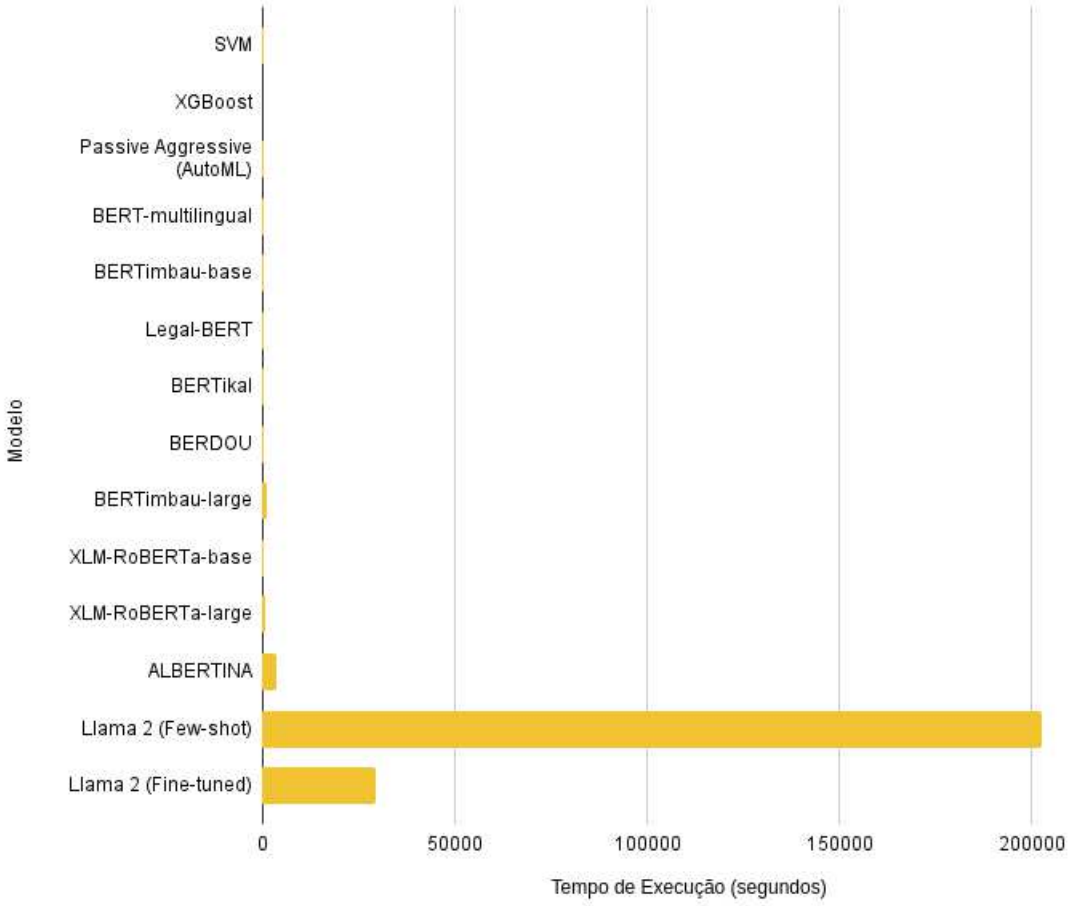
Figura 5.8: Tempo de processamento dos modelos para os conjuntos de treinamento e teste (exceto Llama 2 em *few-shot*).



Fonte: De autoria própria.

dos de teste. Ou seja, para o mesmo conjunto de testes, o modelo BERTimbau-large utilizou aproximadamente 4,9% do tempo de processamento total consumido pelo Llama 2 no cenário desbalanceado. Ao considerar o cenário do Llama 2 na abordagem *few-shot*, foi utilizado apenas 0,52% do tempo pelo BERTimbau-large. As Figuras 5.8 e 5.9 representam grafica-

Figura 5.9: Tempo de processamento dos modelos para o conjunto de teste.



Fonte: De autoria própria.

mente o tempo de processamento de cada um dos modelos, de forma a facilitar a comparação. É perceptível a grande diferença existente entre o Llama 2, o LLM *decoder* utilizado, e os demais modelos, principalmente no cenário *few-shot*.

5.4 Considerações Finais

De maneira geral, os experimentos conduzidos indicaram que os LLMs de arquitetura *encoder* demonstraram resultados superiores em comparação ao LLM de arquitetura *decoder*, obtidos em um período de processamento menor, ou seja, melhores eficácia e eficiência, respectivamente. Além disso, tais LLMs de arquitetura *encoder* demandaram menos recursos de hardware, dispensando a necessidade de otimizações e a aplicação de técnicas de PEFT, tornando-os compatíveis com hardware de uso geral, maximizando a eficiência dos

mesmos. Essas características mostraram-se relevantes nesta pesquisa, exercendo influência decisiva, especialmente na tomada de decisão quanto à adoção desses modelos em um ambiente produtivo, onde é necessário encontrar um equilíbrio entre eficiência e eficácia, ou seja, consumo de recursos e os resultados alcançados.

Entretanto, mesmo que os resultados dos LLMs de arquitetura *encoder* tivessem alcançado métricas superiores aos de arquitetura *decoder*, não significaria um descarte dos modelos menores em um cenário de aplicação real. Dado o alto custo e complexidades adicionais, uma solução híbrida, envolvendo múltiplos modelos, poderia ser uma saída viável. Apesar dos resultados associados ao BERTimbau-large, o modelo mais eficaz, indicarem uma precisão modesta para o limiar de classificação de 0,5 (especificamente, 0,64), a revocação de 0,97 denota um desempenho significativo. No âmbito da análise do DOU, à luz dos valores registrados na matriz de confusão representada na Figura 5.5, a ocorrência de 169 registros Falsos Positivos pode ser considerada uma contrapartida aceitável para identificar com precisão 304 publicações tributárias de maneira automática, dentro de um universo total de 344.569 publicações.

Apesar de grande parte dos registros positivos ter sido capturada, os 11 registros Falsos Negativos não podem ser ignorados. Falsos Negativos representam publicações tributárias que não foram devidamente classificadas pelo modelo, não sendo, portanto, detectadas entre as demais publicações. Isso pode implicar em leis ou alterações tributárias não sendo mapeadas em um fluxo de utilização real. Buscar mapear esses registros é extremamente importante; no entanto, sua ocorrência não diminui a qualidade do resultado encontrado.

No próximo capítulo, são apresentadas as conclusões desta pesquisa, bem como discorrido acerca das ameaças à validade dos resultados, além dos direcionamentos para trabalhos futuros que podem dar continuidade a este trabalho.

Capítulo 6

Conclusão

Os Diários Oficiais desempenham um papel crucial como fontes abrangentes de informações para a sociedade, influenciando tanto o setor público quanto o privado. Suas publicações, que incluem licitações, atos de pessoal, leis e normativos em diversos subdomínios como tributário, trabalhista e ambiental, são caracterizadas por uma linguagem predominantemente jurídica e extensão considerável, especialmente no caso do DOU. Isso torna a análise manual desses documentos custosa, configurando um cenário de aplicação de técnicas PLN, juntamente com o interesse em processar seu conteúdo. Em particular, a área do direito tributário destaca-se como um subdomínio relevante, sendo essencial para empresas manterem o *compliance* com as normas tributárias frequentemente alteradas e publicadas nos Diários Oficiais. O uso de técnicas de PLN nesse contexto contribui para uma identificação eficiente de mudanças tributárias, otimizando recursos humanos, agregando valor e tecnologia ao ramo tributário.

A partir do observado frente ao contexto descrito, juntamente do *corpus* de domínio tributário obtido previamente, o problema-alvo deste trabalho foi direcionado ao contexto do monitoramento de publicações de domínio tributário em meio ao DOU, utilizando-se de técnicas de AM e PLN. Para a condução da pesquisa, um novo conjunto de dados anotado de forma automática foi criado, tendo foco na classificação binária de textos tributários extraídos do DOU. Dentre os modelos de classificação utilizados, foram incluídas técnicas tradicionais de classificação, especificamente o SVM e o XGBoost, além de LLMs, tanto de arquitetura *encoder*, a exemplos do BERTimbau e ALBERTINA, como *decoder*, especificamente o Llama 2.

O modelo de melhor desempenho foi o BERTimbau-large, alcançando um PR-AUC de 0,849973, sendo pelo menos 10,53% superior aos resultados obtidos com os demais modelos utilizados. Com uma revocação de 0,97 e um F1-Score de 0,77 para a classe positiva, o melhor modelo obtido mostrou-se adequado para a tarefa alvo de classificação binária, mesmo tendo como contexto um cenário altamente desbalanceado, com a ocorrência de publicações tributárias sendo apenas aproximadamente 0,09% em relação ao *corpus* total. Grande parte dos modelos tradicionais não alcançou resultados satisfatórios, com PR-AUC em torno de 0,5, exceto para o XGBoost, no cenário balanceado, e o *Passive Aggressive*, no cenário desbalanceado. Quanto ao modelo Llama 2, o LLM de arquitetura *decoder* utilizado nos experimentos, a abordagem de aprendizagem *few-shot* não produziu resultados satisfatórios, atribuindo a classe positiva para todo o conjunto de testes. Após o processo de *fine-tuning* utilizando-se QLoRA para otimização de treinamento e quantização, os resultados do Llama 2 melhoraram em relação à abordagem *few-shot*, mas ainda bem aquém dos resultados dos LLMs *encoder*, resultando em um F1-Score de 0,15 para a classe positiva, em comparação com o 0,77 obtido pelo melhor modelo nos experimentos, o BERTimbau-large.

Os objetivos específicos definidos neste trabalho foram, então, atingidos. A iniciar pelo objetivo específico 1, satisfeito por meio da obtenção de dados de publicações do DOU a partir do *Web Scraper* desenvolvido, tanto para monitoramento de novas publicações como para construção de um *corpus* englobando as publicações disponíveis. Quanto ao objetivo específico 2, foi satisfeito por meio da criação do conjunto de dados anotado automaticamente, utilizando o conjunto de dados original e o *corpus* obtido utilizando-se do *Web Scraper*. Já o objetivo específico 3, o último elencado, foi satisfeito por meio dos diversos experimentos realizados com modelos de classificação, além dos resultados sugerirem que o melhor dos modelos obtidos atende ao problema-alvo deste trabalho.

Esta pesquisa resultou na solução da problemática levantada, bem como também proporcionou diversas contribuições significativas. Inicialmente, pela aplicação de técnicas de AM e PLN de estado da arte de forma pioneira em um domínio ainda pouco explorado na literatura, o domínio jurídico com foco em direito tributário. Outra contribuição é a da aplicação de diversos LLMs num contexto altamente desbalanceado, normalmente configurando em um desafio na resolução de problemas de AM e PLN. Os experimentos evidenciaram que a introdução de desbalanceamento no conjunto de treinamento pode conduzir a resultados

superiores, quando contrastado com um conjunto de treinamento balanceado pela estratégia de *undersampling*.

Adicionalmente, obteve-se a contribuição acerca da aplicação de LLMs pré-treinados em domínios específicos alinhados com o domínio do problema-alvo. Os resultados obtidos representam evidências de que não há garantia de que LLMs pré-treinados em domínios específicos possuem melhor desempenho quando comparados com LLMs de domínio geral, não devendo descartar estes numa análise comparativa e extensiva em busca de melhores resultados. Na tarefa de classificação de publicações tributárias no DOU, observou-se que LLMs pré-treinados, tanto no domínio jurídico como no próprio DOU, não proporcionaram vantagens significativas na obtenção de melhores resultados. Fatores como o desbalanceamento, bem como a proporção entre classes dos dados aplicada ao conjunto de treinamento, podem ter exercido influência determinante no contexto em questão.

Por fim, os resultados obtidos geraram evidências na comparação de LLMs de arquitetura *encoder* e *decoder*, tanto em termos de performance nas métricas obtidas como eficiência no tempo de processamento. Observou-se que, além de resultar em melhores métricas, o modelo BERTimbau-large utilizou apenas 4,9% do tempo de processamento necessário para a abordagem com o Llama 2 ajustado, que ainda necessitou a aplicação de técnicas de quantização e PEFT, como QLoRA, para ser utilizado adequadamente no ambiente computacional disponível. Ao considerar-se a abordagem *few-shot* com o Llama 2, esta proporção torna-se ainda mais discrepante, tendo sido utilizado apenas 0,52% do tempo total pelo BERTimbau-large. Essa constatação indica que, para o problema específico em questão, os LLMs de arquitetura *encoder* ainda são alternativas viáveis, não apenas em termos de resultados, mas também em eficiência de processamento, mesmo com uma menor quantidade de parâmetros em comparação aos LLMs *decoder* de aproximadamente 7 bilhões de parâmetros.

6.1 Questões de Pesquisa

Este trabalho, além da realização dos seus objetivos, buscou responder as questões de pesquisa elencadas no Capítulo 1, as quais revisitamos a seguir:

1. **Os Modelos de Linguagem Grandes (do inglês, *Large Language Models*, ou LLMs) são adequados para a tarefa de classificação de atos de domínio tributário no DOU?**

Com relação à questão de pesquisa 1, verificou-se que o melhor modelo obtido, treinado a partir do BERTimbau-large, é sim adequado para a aplicação proposta. Com um revocação de 0,97, precisão de 0,64 e um F1-Score de 0,77, foi possível identificar 304 dos 315 registros positivos existentes no conjunto de testes, em meio a um total de 344.569 instâncias. Apenas 169 falsos positivos foram apontados pelo modelo. Num cenário de aplicação real, ao invés de realizar a análise direta de mais de 344 mil instâncias, o analista tributário poderia validar todas as publicações classificadas como tributárias pelo modelo, reduzindo assim o conjunto de análise para apenas 473 publicações, referentes às classificações positivas. Outra alternativa é utilizar-se da saída do modelo no formato de probabilidade, definindo diferentes limiares para classificação, deve-se porém haver um balanço, dado que pode não apenas ser capturarada uma quantidade maior de Verdadeiros Positivos, mas aumentar também os Falsos Positivos. Sendo assim, verificou-se que o modelo BERTimbau-large utilizado é capaz de auxiliar significativamente no processo de identificação de publicações tributárias em meio ao DOU, principalmente quando utilizado em conjunto com o *Web Scraper* desenvolvido, de forma a obter os atos tributários no mesmo dia de sua publicação.

2. **Quão bons são os LLMs pré-treinados de domínio geral quando comparados com LLMs pré-treinados em *corpora* do domínio do problema (textos jurídicos e, mais especificamente, documentos do DOU) para a atividade de classificação de texto tributário no DOU?**

Para a questão de pesquisa 2, foram comparados os modelos BERTikal, BERDOU e Legal-BERT com o seu modelo de origem, o BERTimbau-base. Observou-se que a utilização dos modelos pré-treinados em domínios específicos, alinhados com o problema-alvo, sendo estes o jurídico e o do DOU, não resultou na obtenção de melhores resultados para o contexto de classificação de publicações tributárias no DOU. Os modelos de pré-treinamento continuado utilizados obtiveram resultados inferiores ou similares aos do BERTimbau-base, em termos de PR-AUC, corroborando, assim, com a conclusão de que nem sempre o pré-treinamento em domínio específico alinhado com o da tarefa-alvo culmina em melhores resultados.

3. **Quais os impactos das diferentes proporções de balanceamento nos dados de treinamento no contexto altamente desbalanceado da classificação de publicações tributá-**

rias no DOU?

Considerando a questão de pesquisa 3, verificou-se que manter o desbalanceamento no conjunto de treinamento para modelos de classificação pode ser benéfico para cenários extremamente desbalanceados. Nos experimentos realizados, foram obtidos melhores resultados para o conjunto de dados com leve desbalanceamento, sendo este inclinado à classe majoritária, em comparação aos balanceados utilizando a estratégia de *undersampling* da classe negativa. Em contrapartida, em cenários altamente desbalanceados como o observado nesta pesquisa, manter a proporção original dos dados pode levar ao sobreajuste, ou *overfitting*, de modelos para a classe majoritária, sendo necessário experimentação para a identificação da proporção ideal entre as classes para treinamento.

4. **Frente ao grande foco recente nos LLMs de arquitetura *decoder*, os LLMs *encoder* apresentam resultados competitivos, mesmo possuindo menos parâmetros, quando comparados a modelos *decoder* da ordem de 7 bilhões de parâmetros?**

Por fim, tratando-se da questão de pesquisa 4, observou-se que, para a tarefa de classificação de publicações tributárias no DOU, foram obtidos melhores resultados utilizando-se de LLMs *encoder* frente ao Llama 2, modelo de arquitetura *decoder* utilizado nos experimentos. O melhor modelo obtido, o BERTimbau-large, possui aproximadamente 330 milhões de parâmetros, frente aos 7 bilhões do Llama 2, uma quantidade aproximadamente 20 vezes maior. Mesmo assim, o F1-Score obtido com o BERTimbau-large foi de 0,77, frente ao 0,15 obtido com o Llama 2. Além disso, o BERTimbau-large levou 17 minutos para processar o conjunto de testes completo, em comparação com as mais de 8 horas para o Llama 2 ajustado com QLoRA e 56 horas para a abordagem *few-shot*. Os resultados evidenciam que LLMs *encoder* ainda podem representar alternativas competitivas frente aos LLMs *decoder* da ordem de 7 bilhões de parâmetros, principalmente em cenários com limitação de hardware, como foi o contexto desta pesquisa. Assim, sendo necessário aplicar estratégias de otimização de forma a ser possível a sua utilização em *few-shot* e *fine-tuning*, o que pode não ser necessário para os modelos *encoder*.

6.2 Ameaças à Validade

Diante da alta complexidade e relevância do problema endereçado por esta pesquisa, uma análise das ameaças à validade torna-se mandatória, sendo elas as ameaças de construção, interna, externa e de conclusão. O levantamento e estudo detalhado de trabalhos relacionados, bem como a situação deste trabalho frente às linhas de pesquisa e lacunas identificadas na análise dos trabalhos existentes, conforme descrito no Capítulo 3 deste documento, visa mitigar ameaças de construção, dado que foram aplicadas técnicas de estado da arte e pioneiras no problema em questão.

Características como a anotação automática do *corpus* utilizado no estudo e o pequeno conjunto de treinamento, principalmente quando comparado ao conjunto utilizado na avaliação final dos modelos, representam ameaças à validade interna. Entretanto, as métricas e avaliações coletadas dos modelos obtidos, juntamente com o detalhamento do processo metodológico aplicado a esta pesquisa, realizado no Capítulo 4, minimizam esse risco. A validação do *corpus* anotado automaticamente junto ao conjunto de dados original buscou contribuir para evitar a anotação incorreta de registros. Além disso, a natureza desbalanceada do domínio alvo do problema caracteriza a existência de poucos registros positivos, sendo um desafio inerente ao domínio alvo desta pesquisa. Esses fatores também contribuem para a minimização do risco à validade interna.

Dada a clara delimitação do escopo de atuação da pesquisa conduzida, bem como os limites das conclusões obtidas, podem haver ameaças à validade externa. Entretanto, a extensa amostra utilizada como conjunto de testes, composta de 344.569, aliada a ter sido obedecida a ordem temporal dos dados, juntamente do fato de ter sido mantido o desbalanceamento observado no cenário real, foram formas de mitigar a extensibilidade das conclusões obtidas neste trabalho à população-alvo do estudo: as publicações tributárias em meio ao DOU.

Por fim, diversos experimentos foram realizados, originando diferentes cenários e resultados para análise. Métricas adequadas foram selecionadas para a natureza do problema, servindo como ferramentas tanto para medir o desempenho dos modelos quanto para compará-los. Essas características do estudo, junto com a análise estatística das métricas obtidas, buscam mitigar ameaças à validade das conclusões.

6.3 Trabalhos Futuros

Apesar dos resultados obtidos com este trabalho satisfazerem o objetivo proposto e apresentarem diversas contribuições, melhorias podem ser realizadas tanto na metodologia utilizada, como na ampliação do escopo de atuação no problema exposto dos Diários Oficiais. A realização dos trabalhos propostos, além de contribuir para a melhoria deste trabalho, darão continuidade às pesquisas na linha do problema de monitoramento e extração de informações em Diários Oficiais, em domínios além do tributário e do DOU. Dessa forma, seguem abaixo sugestões para trabalhos futuros:

- Aprimorar estratégia de correspondência para aperfeiçoar a qualidade do conjunto de dados, ampliando assim a quantidade de instâncias positivas utilizadas no conjunto final, a partir da utilização de *Word Embeddings* ou outras estratégias de *matching*;
- Realizar a obtenção de publicações de outros Diários Oficiais além do DOU, com o processamento de documentos PDF, utilizando técnicas de Visão Computacional com foco em *Document Object Detection* para análise de layout, segmentação e extração de texto;
- Utilizar LLMs *decoder* da ordem de centenas de bilhões de parâmetros em experimentos adicionais, dada a existência de recursos computacionais para tal, comparando resultados com os LLMs *encoder* utilizados nesta pesquisa;
- Construir *corpus* com foco em outros domínios de classificação além do tributário para o DOU e demais Diários Oficiais; e
- Empregar técnicas de Extração de Informações no âmbito das publicações extraídas, com foco no Reconhecimento de Entidades Nomeadas.

Referências Bibliográficas

ABDALLA, H. I.; AMER, A. A.; RAVANA, S. D. Bow-based neural networks vs. cutting-edge models for single-label text classification. **Neural Computing and Applications**, Springer Science and Business Media LLC, v. 35, n. 27, p. 20103–20116, jul. 2023. ISSN 1433-3058. Disponível em: <http://dx.doi.org/10.1007/s00521-023-08754-z>.

AGUIAR, A. *et al.* Using topic modeling in classification of brazilian lawsuits. In: **Lecture Notes in Computer Science**. Springer International Publishing, 2022. p. 233–242. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_22>.

AKIBA, T. *et al.* **Optuna: A Next-generation Hyperparameter Optimization Framework**. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1907.10902>.

ALAMMAR, J. **The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) — jalammar.github.io**. 2018. <https://jalammar.github.io/illustrated-bert/>. [Acesso em: 23 fev. 2024].

ARAÚJO, G. d. S. *et al.* Natural language processing and social media: a systematic mapping on brazilian leading events. In: **Anais do XX Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2023)**. Sociedade Brasileira de Computação - SBC, 2023. (ENIAC 2023). Disponível em: <http://dx.doi.org/10.5753/eniac.2023.234426>.

ARAUJO, P. H. L. D.; CAMPOS, T. D. Topic modelling brazilian supreme court lawsuits. In: **Frontiers in Artificial Intelligence and Applications**. IOS Press, 2020. Disponível em: <https://doi.org/10.3233/faia200855>.

ARAUJO, P. H. Luz de *et al.* VICTOR: a dataset for Brazilian legal documents classification. In: **Proceedings of the Twelfth Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 1449–1458. ISBN 979-10-95546-34-4. Disponível em: <https://aclanthology.org/2020.lrec-1.181>.

ASH, E.; GUILLOT, M.; HAN, L. Machine extraction of tax laws from legislative texts. In: **Proceedings of the Natural Legal Language Processing Workshop 2021**. Association for Computational Linguistics, 2021. Disponível em: <https://doi.org/10.18653/v1/2021.nllp-1.7>.

BECHO, R. **LIÇÕES DE DIREITO TRIBUTÁRIO**. Saraiva Educação S.A., 2017. ISBN 9788502619661. Disponível em: <https://books.google.com.br/books?id=iD1nDwAAQBAJ>.

BLAGEC, K. *et al.* A critical analysis of metrics used for measuring progress in artificial intelligence. **arXiv preprint arXiv:2008.02577**, 2020.

BOYD, K.; ENG, K. H.; PAGE, C. D. Area under the precision-recall curve: Point estimates and confidence intervals. In: BLOCKEEL, H. *et al.* (Ed.). **Machine Learning and Knowledge Discovery in Databases**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 451–466. ISBN 978-3-642-40994-3.

BROWN, T. B. *et al.* **Language Models are Few-Shot Learners**. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2005.14165>.

CAÇÃO, F. N. *et al.* Tracking environmental policy changes in the brazilian federal official gazette. In: **Lecture Notes in Computer Science**. Springer International Publishing, 2022. p. 256–266. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_24>.

CAVALIERI, A. *et al.* An intelligent system for the categorization of question time official documents of the italian chamber of deputies. **Journal of Information Technology &amp Politics**, Informa UK Limited, v. 20, n. 3, p. 213–234, jun. 2022. Disponível em: <https://doi.org/10.1080/19331681.2022.2082622>.

CERVANTES, J. *et al.* A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, v. 408, p. 189–215, 2020. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231220307153>.

CHEN, H. *et al.* A comparative study of automated legal text classification using random forests and deep learning. **Information Processing Management**, v. 59, n. 2, p. 102798, 2022. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457321002764>.

____. A comparative study of automated legal text classification using random forests and deep learning. **Information Processing &amp Management**, Elsevier BV, v. 59, n. 2, p. 102798, mar. 2022. Disponível em: <https://doi.org/10.1016/j.ipm.2021.102798>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, 2016. (KDD '16). Disponível em: <http://dx.doi.org/10.1145/2939672.2939785>.

CLAVIÉ, B.; ALPHONSUS, M. The unreasonable effectiveness of the baseline: Discussing SVMs in legal text classification. In: **Frontiers in Artificial Intelligence and Applications**. IOS Press, 2021. Disponível em: <https://doi.org/10.3233/faia210317>.

CONNEAU, A. *et al.* Unsupervised cross-lingual representation learning at scale. **CoRR**, abs/1911.02116, 2019. Disponível em: <http://arxiv.org/abs/1911.02116>.

CRAMMER, K. *et al.* Online passive-aggressive algorithms. **Journal of Machine Learning Research**, v. 7, n. 19, p. 551–585, 2006. Disponível em: <http://jmlr.org/papers/v7/crammer06a.html>.

DETTMERS, T. *et al.* **QLoRA: Efficient Finetuning of Quantized LLMs**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2305.14314>.

DEVLIN, J. *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**. Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://doi.org/10.18653/v1/n19-1423>.

DOMINGUES, M. **Language Model in the legal domain in Portuguese**. 2022. <https://huggingface.co/dominguesm/legal-bert-base-cased-ptbr/>.

DWIVEDI, Y. K. *et al.* Opinion paper: "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. **International Journal of Information Management**, Elsevier BV, v. 71, p. 102642, ago. 2023. ISSN 0268-4012. Disponível em: <http://dx.doi.org/10.1016/j.ijinfomgt.2023.102642>.

FERRARIO, A.; NAEGELIN, M. The art of natural language processing: Classical, modern and contemporary approaches to text document classification. **SSRN Electronic Journal**, Elsevier BV, 2020. ISSN 1556-5068. Disponível em: <http://dx.doi.org/10.2139/ssrn.3547887>.

FEURER, M. *et al.* Auto-sklearn 2.0: Hands-free automl via meta-learning. **arXiv:2007.04074 [cs.LG]**, 2020.

____. Efficient and robust automated machine learning. In: **Advances in Neural Information Processing Systems 28 (2015)**. [S.l.: s.n.], 2015. p. 2962–2970.

FOLLONI, A.; SIMM, C. B. Direito tributÁrio, complexidade e anÁlise econÔmica do direito. **Revista Eletrônica do Curso de Direito da UFSM**, Universidad Federal de Santa Maria, v. 11, n. 1, p. 49, jun. 2016. ISSN 1981-3694. Disponível em: <http://dx.doi.org/10.5902/1981369419726>.

FREIRE, D. L. *et al.* Exploratory study of data sampling methods for imbalanced legal text classification. In: BRINGAS, P. G. *et al.* (Ed.). **Hybrid Artificial Intelligent Systems**. Cham: Springer Nature Switzerland, 2023. p. 108–120. ISBN 978-3-031-40725-3.

GOMES, T.; LADEIRA, M. A new conceptual framework for enhancing legal information retrieval at the brazilian superior court of justice. In: **Proceedings of the 12th International Conference on Management of Digital EcoSystems**. ACM, 2020. Disponível em: <https://doi.org/10.1145/3415958.3433087>.

GOUTTE, C.; GAUSSIER, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: **Lecture Notes in Computer Science**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, (Lecture notes in computer science). p. 345–359.

GU, Y. *et al.* **Knowledge Distillation of Large Language Models**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2306.08543>.

GU, Y. H. *et al.* Domain-specific language model pre-training for korean tax law classification. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 10, p. 46342–46353, 2022. Disponível em: <https://doi.org/10.1109/access.2022.3164098>.

GUERRA, F. M.; GUERRA, M. V. C. L. Compliance tributário para redução da litigiosidade fiscal: uma retrospectiva da literatura brasileira recente. **Revista Tributária e de Finanças Públicas**, 2023. ISSN 1518-2711.

HANDI, M. U. *et al.* Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. Institute of Electrical and Electronics Engineers (IEEE), nov. 2023. Disponível em: <http://dx.doi.org/10.36227/techrxiv.23589741>.

HASIB, K. M. *et al.* Strategies for enhancing the performance of news article classification in bangla: Handling imbalance and interpretation. **Engineering Applications of Artificial Intelligence**, Elsevier BV, v. 125, p. 106688, out. 2023. ISSN 0952-1976. Disponível em: <http://dx.doi.org/10.1016/j.engappai.2023.106688>.

HE, P. *et al.* Deberta: decoding-enhanced bert with disentangled attention. In: **9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021**. OpenReview.net, 2021. Disponível em: <https://openreview.net/forum?id=XPZIaotutsD>.

HENDRAWAN, I. R.; UTAMI, E.; HARTANTO, A. D. Comparison of naïve bayes algorithm and xgboost on local product review text classification. **Edumatic: Jurnal Pendidikan Informatika**, v. 6, n. 1, p. 143–149, 2022.

HOW do Transformers work? - Hugging Face NLP Course — huggingface.co. [s.d.]. Disponível em: <https://huggingface.co/learn/nlp-course/en/chapter1/4?fw=pt>. Acesso em: 20 de jan. de 2024.

IMRAN, A. S. *et al.* Classifying european court of human rights cases using transformer-based techniques. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 11, p. 55664–55676, 2023. ISSN 2169-3536. Disponível em: <http://dx.doi.org/10.1109/ACCESS.2023.3279034>.

KAMRAN, M.; SAEED, A.; ALMAGHTHAWI, A. Federated-learning topic modeling based text classification regarding hate speech during covid-19 pandemic. **International Journal of Advanced Computer Science and Applications**, The Science and Information Organization, v. 14, n. 11, 2023. ISSN 2158-107X. Disponível em: <http://dx.doi.org/10.14569/IJACSA.2023.0141157>.

KEILWAGEN, J.; GROSSE, I.; GRAU, J. Area under precision-recall curves for weighted and unweighted data. **PLoS One**, Public Library of Science (PLoS), v. 9, n. 3, p. e92209, mar. 2014.

KHURANA, D. *et al.* Natural language processing: state of the art, current trends and challenges. **Multimedia Tools and Applications**, Springer Science and Business Media LLC, v. 82, n. 3, p. 3713–3744, jul. 2022. ISSN 1573-7721. Disponível em: <http://dx.doi.org/10.1007/s11042-022-13428-4>.

____. Natural language processing: state of the art, current trends and challenges. **Multimedia Tools and Applications**, Springer Science and Business Media LLC, v. 82, n. 3, p. 3713–3744, jul. 2022. ISSN 1573-7721. Disponível em: <http://dx.doi.org/10.1007/s11042-022-13428-4>.

KUMAR, P. N. Detection of textual propaganda using passive aggressive classifiers. **International Journal of Advanced Trends in Computer Science and Engineering**, The World Academy of Research in Science and Engineering, v. 12, n. 2, p. 73–79, abr. 2023. ISSN 2278-3091. Disponível em: <http://dx.doi.org/10.30534/ijatcse/2023/071222023>.

LAURIOLA, I.; LAVELLI, A.; AIOLLI, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. **Neurocomputing**, Elsevier BV, v. 470, p. 443–456, jan. 2022. ISSN 0925-2312. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2021.05.103>.

LEI Nº 5.172, DE 25 DE OUTUBRO DE 1966 - Imprensa Nacional. 1966. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l5172compilado.htm>. Acesso em: 20 de jan. de 2024.

LIU, Y. *et al.* **Understanding LLMs: A Comprehensive Overview from Training to Inference**. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2401.02038>.

LOCKARD, K.; SLATER, R.; SUCRESE, B. Using nlp to model us supreme court cases. **SMU Data Science Review**, v. 7, n. 1, p. 4, 2023.

LOSHCHILOV, I.; HUTTER, F. **Decoupled Weight Decay Regularization**. arXiv, 2017. Disponível em: <https://arxiv.org/abs/1711.05101>.

MACEDO, S. **A importância dos Diários Oficiais - Assembleia Legislativa de Sergipe — al.se.leg.br**. 2018. <https://al.se.leg.br/a-importancia-dos-diarios-oficiais/>. [Acesso em: 2 jan. 2024].

MAH, P. M.; SKALNA, I.; MUZAM, J. Natural language processing and artificial intelligence for enterprise management in the era of industry 4.0. **Applied Sciences**, MDPI AG, v. 12, n. 18, p. 9207, set. 2022. ISSN 2076-3417. Disponível em: <http://dx.doi.org/10.3390/app12189207>.

MAMOOLER, S. *et al.* An efficient active learning pipeline for legal text classification. In: **Proceedings of the Natural Legal Language Processing Workshop 2022**. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 2022. p. 345–358. Disponível em: <https://aclanthology.org/2022.nllp-1.32>.

MANDELBAUM, A.; SHALEV, A. Word embeddings and their use in sentence classification tasks. **arXiv preprint arXiv:1610.08229**, 2016.

MANGRULKAR, S. *et al.* **PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods**. 2022. <https://github.com/huggingface/peft>.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **Introduction to Information Retrieval**. Cambridge, England: Cambridge University Press, 2008.

MARTINS, V. S.; SILVA, C. D. Text classification in law area: a systematic review. In: **Anais do IX Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2021)**. Sociedade Brasileira de Computação - SBC, 2021. Disponível em: <https://doi.org/10.5753/kdmile.2021.17458>.

MENEZES, L. A. M. d.; A, M. A. d. A. G. Do Diario Official do Imperio do Brazil e DiÃ¡rio Oficial da UniÃao e-DiÃ¡rio Oficial: conjunturas e sentidos (1862-2019). **PopulaÃ§Ãe Sociedade**, scielopt, p. 51 – 64, 12 2019. ISSN 2184-5263. Disponível em: <http://scielo.pt/scielo.php?script=sci_arttext&pid=S2184-52632019000200051&nrm=iso>.

NAVEED, H. *et al.* **A Comprehensive Overview of Large Language Models**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2307.06435>.

NGHIEM, M.-Q. *et al.* Text classification and prediction in the legal domain. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2022. p. 4717–4722. Disponível em: <https://aclanthology.org/2022.lrec-1.504>.

NGUYEN, H.-T. *et al.* Transformer-based approaches for legal text processing. **The Review of Socionetwork Strategies**, Springer Science and Business Media LLC, v. 16, n. 1, p. 135–155, jan. 2022. Disponível em: <https://doi.org/10.1007/s12626-022-00102-2>.

NOZZA, D.; BIANCHI, F.; HOVY, D. **What the [MASK]? Making Sense of Language-Specific BERT Models**. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2003.02912>.

PAYING Taxes 2018. Thirteem years of data and analysis on tax systems in 190 economies: A look at recent developments and historical trends. 2018. <https://www.pwc.com/gx/en/paying-taxes/pdf/pwc_paying_taxes_2018_full_report.pdf>. Acesso em: 20 de jan. de 2024].

PAYING Taxes 2020: The changing landscape of tax policy and administration across 190 economies. 2020. <https://www.pwc.com/gx/en/paying-taxes/pdf/pwc-paying-taxes-2020.pdf>. Acesso em: 20 de jan. de 2024].

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PIRES, R. *et al.* Sequence-to-sequence models for extracting information from registration and legal documents. In: **Document Analysis Systems**. Springer International Publishing, 2022. p. 83–95. Disponível em: <https://doi.org/10.1007/978-3-031-06555-2_6>.

POLO, F. M. *et al.* Legalnlp-natural language processing methods for the brazilian legal language. In: SBC. **Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2021. p. 763–774.

PORTARIA IN/SG/PR Nº 9, DE 4 de Fevereiro de 2021 - Imprensa Nacional. 2021. Disponível em: <https://www.in.gov.br/en/web/dou/-/portaria-in/sg/pr-n-9-de-4-de-fevereiro-de-2021-302540550>. Acesso em: 20 de jan. de 2024.

PORTARIA Nº 283, DE 2 DE OUTUBRO DE 2018 - Imprensa Nacional. 2018. Disponível em: <https://in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/43716640/do1-2018-10-03-portaria-n-283-de-2-de-outubro-de-2018-43716563>. Acesso em: 31 de dez. 2023.

PORTO, G. Ensaio sobre os custos de conformidade no brasil: Análise do peso das obrigações tributárias acessórias. **Direito Tributário em Questão**, Universidad Federal de Santa Maria, 2019.

PRIYA, B.; NANDHINI, J. M.; GNANASEKARAN, T. An analysis of the applications of natural language processing in various sectors. In: ____. **Smart Intelligent Computing and Communication Technology**. IOS Press, 2021. Disponível em: <http://dx.doi.org/10.3233/APC210109>.

QI, Q. *et al.* Stochastic optimization of areas under precision-recall curves with provable convergence. **Advances in neural information processing systems**, v. 34, p. 1752–1765, 2021.

QI, Z. The text classification of theft crime based on tf-idf and xgboost model. In: **2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)**. [S.l.: s.n.], 2020. p. 1241–1246.

QUEM Pode Publicar no Diário Oficial da União? | Blog E-DOU — e-dou.com.br. 2022. Disponível em: <https://e-dou.com.br/quem-pode-publicar-no-diario-oficial-da-uniao/>. Acesso em: 2 de jan. 2024.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. OpenAI, 2018. Disponível em: <https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf>.

____. Language models are unsupervised multitask learners. In: . [s.n.], 2019. [Acesso em: 23 fev. 2024]. Disponível em: <https://api.semanticscholar.org/CorpusID:160025533>.

RASCHKA, S. **Finetuning LLMs with LoRA and QLoRA: Insights from Hundreds of Experiments - Lightning AI — lightning.ai**. 2023. <https://lightning.ai/pages/community/lora-insights/>. [Acesso em: 23 fev. 2024].

RENZE, M.; GUVEN, E. **The Effect of Sampling Temperature on Problem Solving in Large Language Models**. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2402.05201>.

ROCHA, L. *et al.* Temporal contexts: Effective text classification in evolving document collections. **Information Systems**, Elsevier BV, v. 38, n. 3, p. 388–409, maio 2013. ISSN 0306-4379. Disponível em: <http://dx.doi.org/10.1016/j.is.2012.11.001>.

____. Exploiting temporal contexts in text classification. In: **Proceedings of the 17th ACM conference on Information and knowledge management**. ACM, 2008. (CIKM08). Disponível em: <http://dx.doi.org/10.1145/1458082.1458117>.

RODRIGUES, J. *et al.* **Advancing Neural Encoding of Portuguese with Transformer Albertina PT-***. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2305.06721>.

RÖNNQVIST, S. *et al.* Is multilingual BERT fluent in language generation? In: NIVRE, J. *et al.* (Ed.). **Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing**. Turku, Finland: Linköping University Electronic Press, 2019. p. 29–36. Disponível em: <https://aclanthology.org/W19-6204>.

SALLES, T. *et al.* A quantitative analysis of the temporal effects on automatic text classification. **Journal of the Association for Information Science and Technology**, Wiley, v. 67, n. 7, p. 1639–1667, ago. 2015. ISSN 2330-1643. Disponível em: <http://dx.doi.org/10.1002/asi.23452>.

SANH, V. *et al.* Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. **CoRR**, abs/1910.01108, 2019. Disponível em: <http://arxiv.org/abs/1910.01108>.

SENGUPTA, S.; DAVE, V. Predicting applicable law sections from judicial case reports using legislative text analysis with machine learning. **Journal of Computational Social Science**, Springer Science and Business Media LLC, v. 5, n. 1, p. 503–516, jul. 2021. Disponível em: <https://doi.org/10.1007/s42001-021-00135-7>.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000.

SILVA, L. **O Que é Diário Oficial e Como Funciona? | e-Diário Oficial — e-diariooficial.com**. 2019. <https://e-diariooficial.com/diario-oficial-o-que-e-e-como-funciona/>. [Acesso em: 2 jan. 2024].

SINGH, V. **Building LLM Applications: Large Language Models (Part 6) — vipra**$_{s}ingh.2024.$**<>**. [Acesso em: 23 fev. 2024].

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT models for brazilian portuguese. In: **Intelligent Systems**. Springer International Publishing, 2020. p. 403–417. Disponível em: <https://doi.org/10.1007/978-3-030-61377-8_28>.

TIWARI, A. Chapter 2 - supervised learning: From theory to applications. In: PANDEY, R. *et al.* (Ed.). **Artificial Intelligence and Machine Learning for EDGE Computing**. Academic Press, 2022. p. 23–32. ISBN 978-0-12-824054-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128240540000265>.

TOUVRON, H. *et al.* **LLaMA: Open and Efficient Foundation Language Models**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2302.13971>.

____. **Llama 2: Open Foundation and Fine-Tuned Chat Models**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2307.09288>.

VASWANI, A. *et al.* Attention is all you need. In: GUYON, I. *et al.* (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

VIRTANEN, A. *et al.* **Multilingual is not enough: BERT for Finnish**. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1912.07076>.

VITALIS, A. Compliance fiscal e regulação fiscal cooperativa. **Revista Direito GV**, FapUNIFESP (SciELO), v. 15, n. 1, 2019. ISSN 2317-6172. Disponível em: <http://dx.doi.org/10.1590/2317-6172201904>.

VRIES, A. de. The growing energy footprint of artificial intelligence. **Joule**, Elsevier BV, v. 7, n. 10, p. 2191–2194, out. 2023. ISSN 2542-4351. Disponível em: <http://dx.doi.org/10.1016/j.joule.2023.09.004>.

WAN, Z. *et al.* **Efficient Large Language Models: A Survey**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2312.03863>.

WANG, C.; LIU, S. X.; AWADALLAH, A. H. **Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference**. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2303.04673>.

WANG, S.; ZHOU, W.; JIANG, C. A survey of word embeddings based on deep learning. **Computing**, Springer Science and Business Media LLC, v. 102, n. 3, p. 717–740, nov. 2019. ISSN 1436-5057. Disponível em: <http://dx.doi.org/10.1007/s00607-019-00768-7>.

XGBOOST - GeeksforGeeks — geeksforgeeks.org. 2023. <https://www.geeksforgeeks.org/xgboost/>. [Acesso em: 23 fev. 2024].

YANG, Z. *et al.* Text classification of judgement documents considering sample imbalance. In: **2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)**. [S.l.: s.n.], 2022. p. 1459–1462.

YU, H. *et al.* **Open, Closed, or Small Language Models for Text Classification?** arXiv, 2023. Disponível em: <https://arxiv.org/abs/2308.10092>.

ZHONG, Q. *et al.* Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. **arXiv preprint**, 2023. Disponível em: <https://arxiv.org/abs/2302.10198>.

Apêndice A

Análise da Obtenção e Formato dos principais Diários Oficiais

De forma a entender como os Diários Oficiais são publicados e o seu formato geral, foi realizado um levantamento com os principais diários de interesse. A análise foi realizada no DOU, nos DOEs dos 26 estados e Distrito Federal, nos DOMs das 26 capitais (o conteúdo referente à Brasília é publicado em conjunto ao do Distrito Federal) e também nos DOMs de mais 23 municípios de interesse, o que totaliza 77 Diários Oficiais. Neste levantamento, em resumo, foram observados os seguintes pontos para cada diário:

- O diário possui PDF completo referente a todo conteúdo publicado por dia?
- O PDF é nativo-digital, ou seja, "pesquisável"?
- O acesso aos arquivos é feito mediante preenchimento de CAPTCHA?
- O diário necessita atenção especial (Seja pela difícil obtenção dos arquivos, pela não disponibilização dos mesmos ou pelos arquivos serem parcialmente ou não "pesquisáveis")?

Além dos pontos acima, também foram levantados pontos quanto à obtenção dos arquivos, para a possível utilização de um crawler para a obtenção dos arquivos, links de acesso e comentários individuais por diário. Estes itens não serão detalhados na Tabela A.1 localizada abaixo, dado que a mesma foi resumida para a apresentação da informação de maior interesse neste momento. Entretanto, a planilha completa poderá ser acessada diretamente

por meio do link compartilhado, esta possui todos os comentários e pontos do estudo realizado. Na Figura A.1 é possível observar um gráfico que destaca a diferença da proporção dos diários que possuem a necessidade de alguma atenção especial frente aos que não.

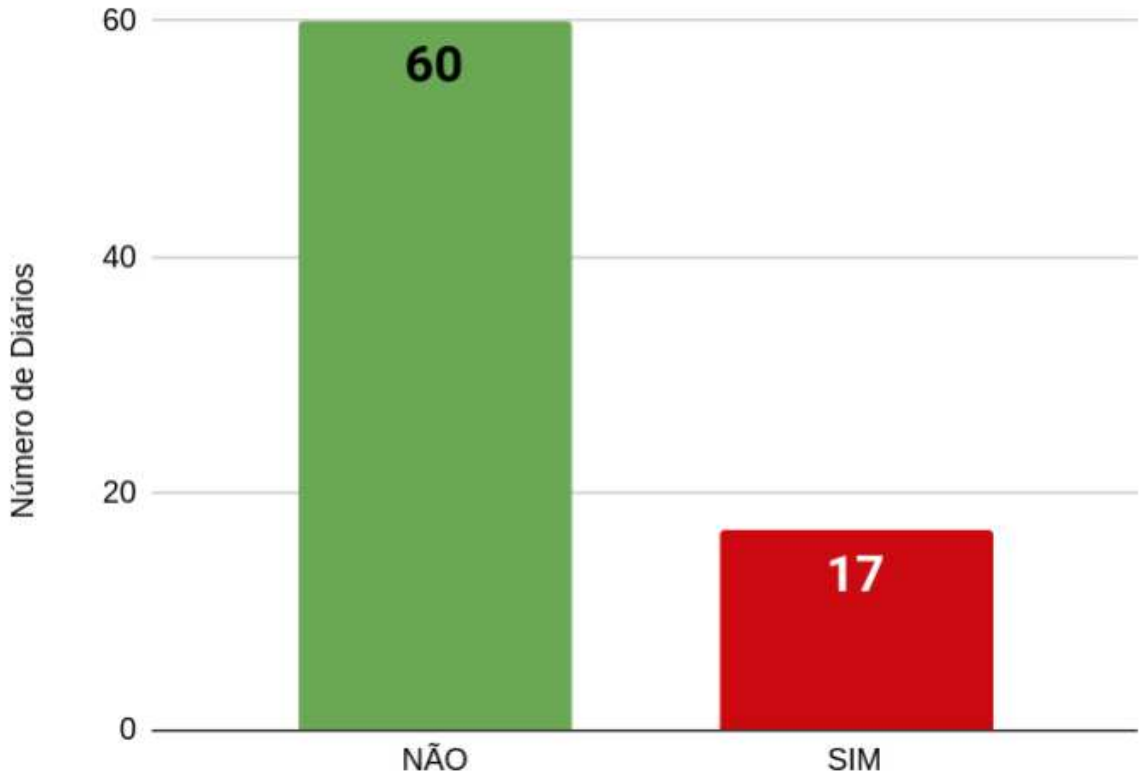


Figura A.1: Quantidade de Diários Oficiais que necessitam de atenção especial.

De forma mais específica, segue abaixo a relação dos Diários Oficiais que necessitam atenção especial, juntamente dos pontos identificados (diários repetidos significa que o mesmo possui mais de um dos problemas relatados) que ocasionaram a presença de cada um nesta lista:

- CAPTCHA:
  - Paraná;
  - Fortaleza;
- PDFs não disponíveis:
  - Bahia (HTML e Imagem);
  - Sergipe (Imagem);

- Formulários:
  - Maranhão;
  - Paraná;
  - Fortaleza;
- PDF página a página (arquivo PDF não é acessível de forma completa):
  - Sergipe;
  - São Paulo (Estado e Capital);
- PDFs não pesquisáveis:
  - Paraíba;
  - Santa Catarina;
  - Aracaju;
  - João Pessoa;
  - Macapá;
  - São Luís;
  - Jaboatão dos Guararapes;
  - Piracicaba;
  - Santo André;
  - Sorocaba.

Legenda das colunas da Tabela A.1:

- *A:* diário possui PDF completo disponível?
- *B:* diário possui PDF pesquisável?
- *C:* acesso ao diário possui CAPTCHA?
- *D:* diário analisado foi classificado como necessitado de atenção especial?
- *E:* data da análise.

Tabela A.1: Análise individual de Diários Oficiais

DIÁRIO	TIPO	A	B	C	D	E
UNIÃO	UNIÃO	SIM	SIM	NÃO	NÃO	26/07/22
ACRE	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
ALAGOAS	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
AMAPÁ	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
AMAZONAS	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
BAHIA	ESTADO	NÃO	NÃO	NÃO	SIM	26/07/22
CEARÁ	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
DISTRITO FEDERAL	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
ESPIRITO SANTO	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
GOIÁS	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
MARANHÃO	ESTADO	NÃO	SIM	NÃO	SIM	26/07/22
MATO GROSSO	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
MATO GROSSO DO SUL	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
MINAS GERAIS	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
PARÁ	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
PARAÍBA	ESTADO	SIM	NÃO	NÃO	SIM	26/07/22
PARANÁ	ESTADO	SIM	SIM	SIM	SIM	26/07/22
PERNAMBUCO	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
PIAUÍ	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
RIO DE JANEIRO	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
RIO GRANDE DO NORTE	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22

Continuado na próxima página

Tabela A.1: Análise individual de Diários Oficiais (Continuado)

RIO GRANDE DO SUL	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
RONDÔNIA	ESTADO	SIM	SIM	NÃO	NÃO	26/07/22
RORAIMA	ESTADO	SIM	SIM	NÃO	NÃO	28/07/22
SANTA CATARINA	ESTADO	SIM	NÃO	NÃO	SIM	28/07/22
SÃO PAULO	ESTADO	NÃO	SIM	NÃO	SIM	28/07/22
SERGIPE	ESTADO	NÃO	NÃO	NÃO	SIM	28/07/22
TOCANTINS	ESTADO	SIM	SIM	NÃO	NÃO	28/07/22
ARACAJU	CAPITAL	SIM	NÃO	NÃO	SIM	28/07/22
BELÉM	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
BELO HORIZONTE	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
BOA VISTA	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
CAMPO GRANDE	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
CUIABÁ	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
CURITIBA	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
FLORIANÓPOLIS	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
FORTALEZA	CAPITAL	SIM	SIM	SIM	SIM	28/07/22
GOIÂNIA	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
JOÃO PESSOA	CAPITAL	SIM	NÃO	NÃO	SIM	28/07/22
MACAPÁ	CAPITAL	SIM	NÃO	NÃO	SIM	28/07/22
MACEIÓ	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
MANAUS	CAPITAL	SIM	SIM	NÃO	NÃO	28/07/22
NATAL	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
PALMAS	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22

Continuado na próxima página

Tabela A.1: Análise individual de Diários Oficiais (Continuado)

PORTO ALEGRE	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
PORTO VELHO	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
RECIFE	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
RIO BRANCO	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
RIO DE JANEIRO	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
SALVADOR	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
SÃO LUIZ	CAPITAL	SIM	NÃO	NÃO	SIM	31/07/22
SÃO PAULO	CAPITAL	NÃO	SIM	NÃO	SIM	31/07/22
TERESINA	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
VITÓRIA	CAPITAL	SIM	SIM	NÃO	NÃO	31/07/22
BARUERI	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
CAMPINAS	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
DIADEMA	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
FOZ DO IGUAÇU	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
GUARUJÁ	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
GUARULHOS	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
INDAIATUBA	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
JABOATÃO DOS GUARARAPES	MUNICÍPIO	SIM	NÃO	NÃO	SIM	22/02/23
MARINGÁ	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
NITERÓI	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
OSASCO	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
PAULINIA	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
PETROLINA	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22

Continuado na próxima página

Tabela A.1: Análise individual de Diários Oficiais (Continuado)

PIRACICABA	MUNICÍPIO	SIM	NÃO	NÃO	SIM	31/07/22
RIBEIRÃO PRETO	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
SANTO ANDRÉ	MUNICÍPIO	SIM	NÃO	NÃO	SIM	31/07/22
SANTOS	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
SÃO BERNARDO DOS CAMPOS	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
SÃO CARLOS	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
SÃO JOSÉ DO RIO PRETO	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
SERRA	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22
SOROCABA	MUNICÍPIO	SIM	NÃO	NÃO	SIM	31/07/22
TAUBATÉ	MUNICÍPIO	SIM	SIM	NÃO	NÃO	31/07/22

Apêndice B

Exemplos de Divergência de Layout e Formatação dentre Diários Oficiais

De forma a detalhar os diferentes formatos adotados pelos diversos Diários Oficiais, seguem abaixo diferentes exemplos de Diários Oficiais de relevância. Nos exemplos abaixo será possível identificar as claras diferenças de formatação dos documentos, layout e organização das informações publicadas.

ESTADO DO ACRE
SECRETARIA DE ESTADO DE HABITAÇÃO E URBANISMO – SEHURB

PORTARIA SEHURB Nº 42, DE 15 DE FEVEREIRO DE 2023
O SECRETÁRIO DE ESTADO DE HABITAÇÃO E URBANISMO - SEHURB, em exercício, no uso das atribuições legais que lhe confere o Decreto nº 1.852-P, de 14 de fevereiro de 2023, publicado no Diário Oficial do Estado nº 13.476, de 15 de fevereiro de 2023,
RESOLVE:
Art. 1º Conceder a Função de Confiança do Poder Executivo – FCPE-11 ao Gestor de Políticas Públicas, Alexandre Silva Meireles, matrícula nº 9269975, para responder pela Divisão de Planejamento e pela Divisão de Contabilidade Setorial, cumulativamente, no âmbito da Secretaria de Estado de Habitação e Urbanismo, até ulterior deliberação.
Art. 2º Esta Portaria entra em vigor na data de sua publicação, com efeitos a contar de 2 de janeiro de 2023.

Roberto Derze Craveiro
Secretário de Estado de Habitação e Urbanismo, em exercício
Decreto nº 1.852-P, de 14 de fevereiro de 2023

ESTADO DO ACRE
SECRETARIA DE ESTADO DE HABITAÇÃO E URBANISMO – SEHURB

PORTARIA SEHURB Nº 43, DE 15 DE FEVEREIRO DE 2023
O SECRETÁRIO DE ESTADO DE HABITAÇÃO E URBANISMO - SEHURB, em exercício, no uso das atribuições legais que lhe confere o Decreto nº 1.852-P, de 14 de fevereiro de 2023, publicado no Diário Oficial do Estado nº 13.476, de 15 de fevereiro de 2023,
RESOLVE:
Art. 1º Conceder a Função de Confiança do Poder Executivo – FCPE-11 a Engenheira Civil, Aline Louise Silva Ramos, matrícula nº 9336753, para responder pela Divisão de Convênio e pela Divisão de Gestão e Fiscalização de Contratos, cumulativamente, no âmbito da Secretaria de Estado de Habitação e Urbanismo, até ulterior deliberação.
Art. 2º Esta Portaria entra em vigor na data de sua publicação, com efeitos a contar de 2 de janeiro de 2023.

Roberto Derze Craveiro
Secretário de Estado de Habitação e Urbanismo, em exercício
Decreto nº 1.852-P, de 14 de fevereiro de 2023

ESTADO DO ACRE
SECRETARIA DE ESTADO DE HABITAÇÃO E URBANISMO – SEHURB

PORTARIA SEHURB Nº 44, DE 15 DE FEVEREIRO DE 2023
O SECRETÁRIO DE ESTADO DE HABITAÇÃO E URBANISMO - SEHURB, em exercício, no uso das atribuições legais que lhe confere o Decreto nº 1.852-P, de 14 de fevereiro de 2023, publicado no Diário Oficial do Estado nº 13.476, de 15 de fevereiro de 2023,
RESOLVE:
Art. 1º Conceder a Função de Confiança do Poder Executivo – FCPE-11 a Engenheira Civil, Jessica Laurenti, matrícula nº 9259597, para responder pelo Departamento Administrativo, no âmbito da Secretaria de Estado de Habitação e Urbanismo, até ulterior deliberação.
Art. 2º Esta Portaria entra em vigor na data de sua publicação, com efeitos a contar de 2 de janeiro de 2023.

Roberto Derze Craveiro
Secretário de Estado de Habitação e Urbanismo, em exercício
Decreto nº 1.852-P, de 14 de fevereiro de 2023

ESTADO DO ACRE
SECRETARIA DE ESTADO DE HABITAÇÃO E URBANISMO – SEHURB

PORTARIA SEHURB Nº 45, DE 15 DE FEVEREIRO DE 2023
O SECRETÁRIO DE ESTADO DE HABITAÇÃO E URBANISMO - SEHURB, em exercício, no uso das atribuições legais que lhe confere o Decreto nº 1.852-P, de 14 de fevereiro de 2023, publicado no Diário Oficial do Estado nº 13.476, de 15 de fevereiro de 2023,
RESOLVE:
Art. 1º Conceder a Função de Confiança do Poder Executivo – FCPE-11 a Especialista Executiva, Dayana Silva Araújo, matrícula nº 9345230, para responder pelo Departamento Social, no âmbito da Secretaria de Estado de Habitação e Urbanismo, até ulterior deliberação.
Art. 2º Esta Portaria entra em vigor na data de sua publicação, com efeitos a contar de 2 de janeiro de 2023.

Roberto Derze Craveiro
Secretário de Estado de Habitação e Urbanismo, em exercício
Decreto nº 1.852-P, de 14 de fevereiro de 2023

ESTADO DO ACRE
SECRETARIA DE ESTADO DE HABITAÇÃO E URBANISMO – SEHURB

PORTARIA SEHURB Nº 46, DE 15 DE FEVEREIRO DE 2023
O SECRETÁRIO DE ESTADO DE HABITAÇÃO E URBANISMO - SEHURB, em exercício, no uso das atribuições legais que lhe confere o Decreto nº 1.852-P, de 14 de fevereiro de 2023, publicado no Diário Oficial do Estado nº 13.476, de 15 de fevereiro de 2023,
RESOLVE:
Art. 1º Conceder a Função de Confiança do Poder Executivo – FCPE-11 ao Arquiteto e Urbanista, Leonardo Neder de Faro Freire, matrícula nº 9126953, para responder pelo Departamento de Habitação, no âmbito da Secretaria de Estado de Habitação e Urbanismo, até ulterior deliberação.
Art. 2º Esta Portaria entra em vigor na data de sua publicação, com efeitos a contar de 2 de janeiro de 2023.

Roberto Derze Craveiro
Secretário de Estado de Habitação e Urbanismo, em exercício
Decreto nº 1.852-P, de 14 de fevereiro de 2023

ESTADO DO ACRE
SECRETARIA DE ESTADO DE HABITAÇÃO E URBANISMO – SEHURB

PORTARIA SEHURB Nº 47, DE 15 DE FEVEREIRO DE 2023
O SECRETÁRIO DE ESTADO DE HABITAÇÃO E URBANISMO - SEHURB, em exercício, no uso das atribuições legais que lhe confere o Decreto nº 1.852-P, de 14 de fevereiro de 2023, publicado no Diário Oficial do Estado nº 13.476, de 15 de fevereiro de 2023,
RESOLVE:
Art. 1º Conceder a Função de Confiança do Poder Executivo – FCPE-11 ao Técnico em Contabilidade, Jair Roberto Guedes Gutierrez, matrícula nº 52175, para responder pelo Departamento de Licitações e Contratos, no âmbito da Secretaria de Estado de Habitação e Urbanismo, até ulterior deliberação.
Art. 2º Esta Portaria entra em vigor na data de sua publicação, com efeitos a contar de 2 de janeiro de 2023.

Roberto Derze Craveiro
Secretário de Estado de Habitação e Urbanismo, em exercício
Decreto nº 1.852-P, de 14 de fevereiro de 2023

SEJUSP

PORTARIA SEJUSP Nº 151, DE 14 DE FEVEREIRO DE 2023
PROCESSO SEI Nº : 0819.012827.00002/2023-16
O SECRETÁRIO DE ESTADO DE JUSTIÇA E SEGURANÇA PÚBLICA, JOSÉ AMÉRICO DE SOUZA GAIA, no uso das atribuições que lhe são conferidas, por meio do Decreto nº. 10-P de 01 de janeiro de 2023, publicado no Diário Oficial do Estado nº 13.443, de 02 de janeiro de 2023, em consonância com o Artigo 86, Inciso II, da Constituição do Estado do Acre, de 03 de outubro de 1989;
CONSIDERANDO o teor do DECRETO Nº 1.616-P, DE 6 DE FEVEREIRO DE 2023 (Evento SEI nº 6162715);
CONSIDERANTO o teor do Despacho nº 176/2023/SEJUSP - DAGS (Evento-SEI nº 6166344).
RESOLVE:
Art. 1º - Lotar o servidor 3° SGT PM - EUDALEX DOS SANTOS MELO NASCIMENTO, matrícula nº 9293159-1, na Divisão de Transporte e Segurança Interna - DIVTSI/SEJUSP.
Art. 2º - Esta portaria entra em vigor na data da sua assinatura.
Publique-se e Cumpra-se.

JOSÉ AMÉRICO DE SOUZA GAIA
Secretário de Estado de Justiça e Segurança Pública

PORTARIA SEJUSP Nº 152, DE 15 DE FEVEREIRO DE 2023
PROCESSO SEI Nº : 0819.012828.00007/2023-30
O SECRETÁRIO DE ESTADO DE JUSTIÇA E SEGURANÇA PÚBLICA, JOSÉ AMÉRICO DE SOUZA GAIA, no uso das atribuições que lhe são conferidas, por meio do Decreto nº. 10-P de 01 de janeiro de 2023, publicado no Diário Oficial do Estado nº 13.443, de 02 de janeiro de 2023, em consonância com o Artigo 86, Inciso II, da Constituição do Estado do Acre, de 03 de outubro de 1989;
RESOLVE:
Art. 1º - Lotar a servidora GLEICE PEREIRA JUSTA DA SILVA, matrícula nº 2756676-4, no Controle Interno da Secretaria de Estado de Justiça e Segurança Pública - SEJUSP, e cumulativamente no Controle Interno do Fundo Estadual de Segurança Pública – FUNDESEG;
Art. 2º - Esta portaria entra em vigor na data da sua assinatura.
Publique-se e Cumpra-se.

JOSÉ AMÉRICO DE SOUZA GAIA
Secretário de Estado de Justiça e Segurança Pública

Figura B.1: Exemplo de página com coluna dupla e disposição dos atos (Diário Oficial do Estado do Acre).

16 João Pessoa - Sábado, 18 de Fevereiro de 2023 Diário Oficial

convocada a empresa acima mencionada para a assinatura do contrato.

João Pessoa, 15 de fevereiro de 2023.

Luiz Gustavo César de Barros Correia
Diretor Superintendente

*dados anonimizados.

FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE
DIRETORIA ADMINISTRATIVA E FINANCEIRA
GERÊNCIA EXECUTIVA DE COMPRAS E CONTRATOS

TERMO DE HOMOLOGAÇÃO
E DIVULGAÇÃO DO RESULTADO
PROCESSO Nº PBS-PRC-2022/01029
DISPENSA DE SELEÇÃO DE FORNECEDORES
(art. 37, II do Regulamento Próprio de Compras e Contratações de Serviços)
REGISTRO CGE Nº 23-00216-6

OBJETO: PROCESSO PARA AQUISIÇÃO DE MATERIAL INFORMÁTICO PARA O SETOR DE HEMODINÂMICA DO COMPLEXO HOSPITALAR REGIONAL DEPUTADO JANDUHY CARNEIRO. PATOS / PB - EQUIPAMENTO DESKTOP PARA MELHORIA DA COMUNICAÇÃO ENTRE O HOSPITAL METROPOLITANO DOM JOSÉ MARIA PIRES E O DISPOSITIVO INSTALADO NO SETOR DE HEMODINÂMICA DO COMPLEXO HOSPITALAR REGIONAL DEPUTADO JANDUHY CARNEIRO, BEM COMO PARA EXECUÇÃO INTEGRADA DO SISTEMA DE INFORMAÇÕES RADIOLÓGICAS (RIS), DE MODO À ATENDER ÀS NECESSIDADES DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE.

O DIRETOR SUPERINTENDENTE DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE, com base no Parecer Jurídico nº **0047/2023** – AEAJ e demais peças do processo, em cumprimento ao art. 36 do Regulamento Interno de Compras e Contratações de Serviços (RICCS), **HOMOLOGA E DIVULGA** o resultado da dispensa de seleção de fornecedores em favor da empresa: **CONECT RS SOLUÇÕES EM TECNOLOGIA LTDA,** inscrita no CNPJ sob o nº 48.322.818/0001-57, no valor total R$ 5.599,99 (Cinco Mil e Quinhentos e Noventa e Nove Reais e Noventa e Nove Centavos). Ante o exposto, com fundamento no art. 37, do RICCS ficam convocadas as empresas acima mencionadas para a assinatura do contrato.

João Pessoa, 15 de fevereiro de 2023.

Luiz Gustavo César de Barros Correia
Diretor Superintendente

FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE
DIRETORIA ADMINISTRATIVA E FINANCEIRA
GERÊNCIA EXECUTIVA DE COMPRAS E CONTRATOS

TERMO DE HOMOLOGAÇÃO
E DIVULGAÇÃO DO RESULTADO
PROCESSO Nº PBS-PRC-2022/01031
DISPENSA DE SELEÇÃO DE FORNECEDORES
(art. 37, II do Regulamento Próprio de Compras e Contratações de Serviços)
REGISTRO CGE Nº 23-00219-1

OBJETO: PROCESSO PARA AQUISIÇÃO DE MATERIAIS INFORMÁTICOS PARA O SETOR DE HEMODINÂMICA DO COMPLEXO HOSPITALAR REGIONAL DEPUTADO JANDUHY CARNEIRO. PATOS / PB - APLIANCE FIREWALL PFSENSE, MONITOR DE 17 POLEGADAS LCD OU SUPERIOR, BIVOLT COM FONTE INTERNA E WEBCAM HD 1080P COM MICROFONE EMBUTIDO, DE MODO A ATENDER ÀS NECESSIDADES DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE.

O DIRETOR SUPERINTENDENTE DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE, com base no Parecer Jurídico nº **0051/2023** – AEAJ e demais peças do processo, em cumprimento ao art. 36 do Regulamento Interno de Compras e Contratações de Serviços (RICCS), **HOMOLOGA E DIVULGA** o resultado da dispensa de seleção de fornecedores em favor das empresas: **MONTEIRO SERVICOS E SOLUCOES DE TI LTDA,** inscrita no CNPJ sob o nº 42.559.662/0001-46, no valor total R$ 5.000,00 (Cinco mil Reais); **EXECUTIVE INFORMATICA E SERVICOS LTDA,** inscrita no CNPJ sob o nº 08.309.659/0001-36, no valor total R$ 430,00 (Quatrocentos e Trinta Reais) e **CONECT RS SOLUÇÕES EM TECNOLOGIA LTDA,** inscrita no CNPJ sob o nº 48.322.818/0001-57, no valor total R$ 862,87 (Oitocentos e Sessenta e Dois Reais e Oitenta e Sete Centavos). Perfazendo o total de R$ 6.292,87 (Seis Mil e Duzentos e Noventa e Dois Reais e Oitenta e Sete Centavos). Ante o exposto, com fundamento no art. 37, do RICCS ficam convocadas as empresas acima mencionadas para a assinatura do contrato.

João Pessoa, 15 de fevereiro de 2023.

Luiz Gustavo César de Barros Correia
Diretor Superintendente

FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE
DIRETORIA ADMINISTRATIVA E FINANCEIRA
GERÊNCIA EXECUTIVA DE COMPRAS E CONTRATOS

TERMO DE HOMOLOGAÇÃO
E DIVULGAÇÃO DO RESULTADO
PROCESSO Nº PBS-PRC-2022/00873
DISPENSA DE SELEÇÃO DE FORNECEDORES
(art. 37, II do Regulamento Próprio de Compras e Contratações de Serviços)
REGISTRO CGE Nº 23-00214-0

OBJETO: PROCESSO PARA AQUISIÇÃO E SENSORES DE FLUXO NEONATOS E PEDIÁTRICOS PARA VENTILADORES IX5 INTERMED, DE MODO A ATENDER AS NECESSIDADES DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE.

O DIRETOR SUPERINTENDENTE DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE, com base no Parecer Jurídico nº **0058/2023** – AEAJ e demais peças do processo, em cumprimento ao art. 36 do Regulamento Interno de Compras e Contratações de Serviços (RICCS), **HOMOLOGA E DIVULGA** o resultado da dispensa de seleção de fornecedores em favor da empresa **HBL - VENDAS E SERVICOS DE ARTIGOS MEDICOS E ORTOPEDICOS LTDA,** inscrita no CNPJ sob o nº 05.000.571/0001-40, no valor total de: 2.620,00 (Dois Mil, Seiscentos e Vinte Reais). Ante o exposto, com fundamento no art. 37, do RICCS fica convocada a empresa acima mencionada para a assinatura do contrato.

João Pessoa, 17 de fevereiro de 2023.

Luiz Gustavo César de Barros Correia
Diretor Superintendente

FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE
DIRETORIA ADMINISTRATIVA E FINANCEIRA
GERÊNCIA EXECUTIVA DE COMPRAS E CONTRATOS

TERMO DE HOMOLOGAÇÃO
E DIVULGAÇÃO DO RESULTADO
PROCESSO Nº PBS-PRC-2022/00905
DISPENSA DE SELEÇÃO DE FORNECEDORES
(art. 37, II do Regulamento Próprio de Compras e Contratações de Serviços)
REGISTRO CGE Nº 23-00215-8

OBJETO: PROCESSO PARA AQUISIÇÃO DE MATERIAIS INFORMÁTICOS PARA O SETOR DE HEMODINÂMICA DO COMPLEXO HOSPITALAR REGIONAL DEPUTADO JANDUHY CARNEIRO. PATOS / PB - APLIANCE FIREWALL PFSENSE, MONITOR DE 17 POLEGADAS LCD OU SUPERIOR, BIVOLT COM FONTE INTERNA E WEBCAM HD 1080P COM MICROFONE EMBUTIDO, DE MODO A ATENDER ÀS NECESSIDADES DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE.

O DIRETOR SUPERINTENDENTE DA FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE, com base no Parecer Jurídico nº **0052/2023** – AEAJ e demais peças do processo, em cumprimento ao art. 36 do Regulamento Interno de Compras e Contratações de Serviços (RICCS), **HOMOLOGA E DIVULGA** o resultado da dispensa de seleção de fornecedores em favor das empresas: **PAPELARIA E LIVRARIA PEDRO II LTDA,** inscrita no CNPJ sob o nº 24.116.337/0001-27, no valor total R$ 1.630,00 (Um mil, Seiscentos e Trinta Reais) e **S D COMÉRCIO DE ARTIGOS DE BRINDES E SERVIÇOS GRÁFICOS LTDA,** inscrita no CNPJ sob o nº 41.570.283/0001-94, no valor total R$ 6.750,00 (Seis Mil, Setecentos e Cinquenta Reais). Perfazendo o total de R$ 8.380,00 (Oito Mil, Trezentos e Oitenta Reais). Ante o exposto, com fundamento no art. 37, do RICCS ficam convocadas as empresas acima mencionadas para a assinatura do contrato.

João Pessoa, 15 de fevereiro de 2023.

Luiz Gustavo César de Barros Correia
Diretor Superintendente

FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE

AVISO DE SESSÃO PÚBLICA
PROCESSO Nº PBS-PRC-2022/00669
SELEÇÃO DE FORNECEDORES Nº 02/2023
REGISTRO CGE Nº 23-00213-2
LICITAÇÃO BB 981513

DATA DE ABERTURA DAS PROPOSTAS: 06/03/2023 - às 14h.
INÍCIO DA DISPUTA: 06/03/2023 - às 14h15min.
OBJETO: Aquisição de Kit com Introdutor de Catéter de Eletrodo Bipolar para Marcapasso Temporário 6frx100cm e 5frx100cm e Kit TCA 2000

A **FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE – PB SAÚDE**, Fundação Pública de Direito Privado, por meio de sua Agente de Contratação, Marília Quirino de Almeida, designada pela Portaria nº 0037/2022, torna público, para conhecimento dos interessados, que fará procedimento de Seleção de Fornecedores, na modalidade Pregão do tipo Eletrônico sob o critério de menor preço, nos termos do Regulamento Interno de Compras e Contratações de Serviços (RICCS).

O Edital ficará à disposição dos interessados no prazo prescrito na legislação pertinente no portal da PB SAÚDE através do link https://pbsaude.pb.gov.br/regulamento-proprio/editais-para-a-selecao-de-fornecedores ou no endereço eletrônico do portal www.licitacoes-e.com.br.

Em caso de dúvidas, consultas com a Agente de Contratação no horário das 8h às 12h e das 13h às 16h30min, nos telefones: (83) 3229-9100 e 3229-9576, ou pelo e-mail: selecaodefornecedores.pbsaude@gmail.com

João Pessoa, 17 de fevereiro de 2023.

Marília Quirino de Almeida
Matrícula nº 000021
Agente de Contratação

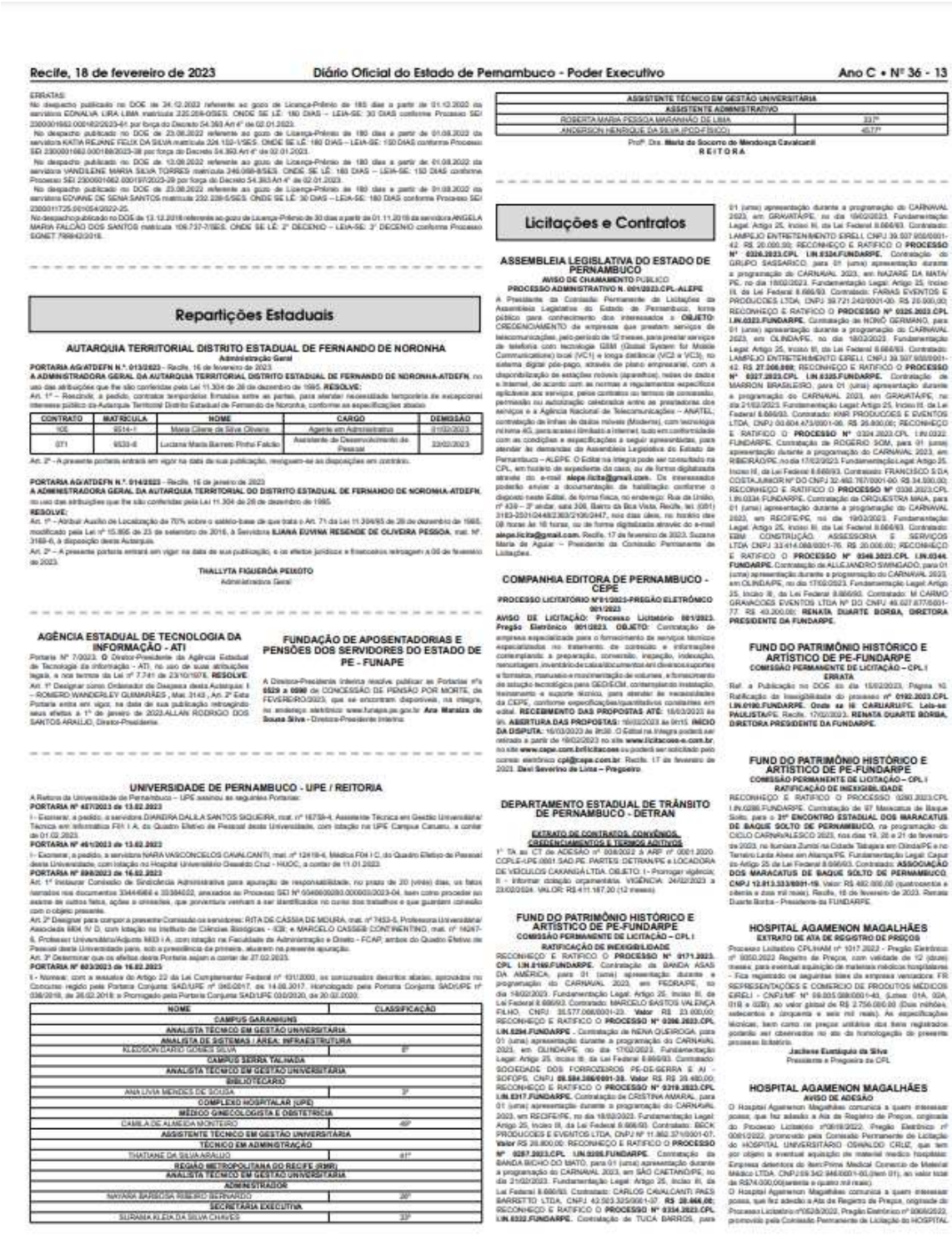
FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE

AVISO DE SESSÃO PÚBLICA (REPUBLICAÇÃO)
PROCESSO Nº PBS-PRC-2022/00920
SELEÇÃO DE FORNECEDORES N 053/2022
REGISTRO CGE Nº 22-02557-2LICITAÇÃO BB 987057

DATA DE ABERTURA DAS PROPOSTAS: 09/03/2023 - às 09h.
INÍCIO DA DISPUTA: 09/03/2023 - às 09h15min
OBJETO: Aquisição de Insumos de Almoxarifado Geral

A **FUNDAÇÃO PARAIBANA DE GESTÃO EM SAÚDE – PB SAÚDE**, Fundação Pública de Direito

Figura B.2: Exemplo de página com coluna dupla e disposição dos atos (Diário Oficial do Estado da Paraíba).



Recife, 18 de fevereiro de 2023 Diário Oficial do Estado de Pernambuco - Poder Executivo Ano C • Nº 36 - 13

ERRATAS:
No despacho publicado no DOE de 24.12.2022 referente ao gozo de Licença-Prêmio de 180 dias a partir de 01.12.2022 da servidora EDNALVA LIRA LIMA matrícula 220.269-0/SES, ONDE SE LÊ: 180 DIAS – LEIA-SE: 30 DIAS conforme Processo SEI 2300001662.000162/2023-61 por força do Decreto 54.393 Art 4º de 02.01.2023.
No despacho publicado no DOE de 23.08.2022 referente ao gozo de Licença-Prêmio de 180 dias a partir de 01.08.2022 da servidora KATIA REJANE FELIX DA SILVA matrícula 224.102-1/SES, ONDE SE LÊ: 180 DIAS – LEIA-SE: 150 DIAS conforme Processo SEI 2300001662.000188/2023-38 por força do Decreto 54.393 Art 4º de 02.01.2023.
No despacho publicado no DOE de 13.08.2022 referente ao gozo de Licença-Prêmio de 180 dias a partir de 01.08.2022 da servidora VANDILENE MARIA SILVA TORRES matrícula 346.068-8/SES, ONDE SE LÊ: 180 DIAS – LEIA-SE: 150 DIAS conforme Processo SEI 2300001662.000197/2023-29 por força do Decreto 54.393 Art 4º de 02.01.2023.
No despacho publicado no DOE de 23.08.2022 referente ao gozo de Licença-Prêmio de 180 dias a partir de 01.08.2022 da servidora EDVANE DE SENA SANTOS matrícula 232.228-5/SES, ONDE SE LÊ: 30 DIAS – LEIA-SE: 180 DIAS conforme Processo SEI 2300011725.001054/2022-25.
No despacho publicado no DOS de 13.12.2018 referente ao gozo de Licença-Prêmio de 30 dias a partir de 01.11.2018 da servidora ANGELA MARIA FALCÃO DOS SANTOS matrícula 106.737-7/SES, ONDE SE LÊ: 2º DECÊNIO – LEIA-SE: 3º DECÊNIO conforme Processo SGNET 76884/2018.

Repartições Estaduais

AUTARQUIA TERRITORIAL DISTRITO ESTADUAL DE FERNANDO DE NORONHA
Administração Geral

PORTARIA AGVATDEFN N.º. 013/2023 - Recife, 16 de fevereiro de 2023
A ADMINISTRADORA GERAL DA AUTARQUIA TERRITORIAL DISTRITO ESTADUAL DE FERNANDO DE NORONHA-ATDEFN, no uso das atribuições que lhe são conferidas pela Lei 11.304 de 28 de dezembro de 1995. **RESOLVE:**
Art. 1º – Rescindir, a pedido, contratos temporários firmados entre as partes, para atender necessidade temporária de excepcional interesse público da Autarquia Territorial Distrito Estadual de Fernando de Noronha, conforme as especificações abaixo:

CONTRATO	MATRÍCULA	NOME	CARGO	DEMISSÃO
10E	9514-1	Maria Cilene da Silva Oliveira	Agente em Administrativo	01/02/2023
07I	9533-8	Luciana Maria Barreto Pinho Falcão	Assistente de Desenvolvimento de Pessoal	23/02/2023

Art. 2º - A presente portaria entrará em vigor na data de sua publicação, revogam-se as disposições em contrário.

PORTARIA AGVATDEFN N.º. 014/2023 - Recife, 16 de janeiro de 2023
A ADMINISTRADORA GERAL DA AUTARQUIA TERRITORIAL DO DISTRITO ESTADUAL DE FERNANDO DE NORONHA-ATDEFN, no uso das atribuições que lhe são conferidas pela Lei 11.304 de 28 de dezembro de 1995.
RESOLVE:
Art. 1º - Atribuir Auxílio de Localização de 70% sobre o salário-base de que trata o Art. 71 da Lei 11.304/95 de 28 de dezembro de 1995, modificado pela Lei nº 15.895 de 23 de setembro de 2016, à Servidora **ELIANA EUVINA RESENDE DE OLIVEIRA PESSOA**, mat. Nº. 3168-6, à disposição desta Autarquia.
Art. 2º – A presente portaria entrará em vigor na data de sua publicação, e os efeitos jurídicos e financeiros retroagem a 09 de fevereiro de 2023.

THALLYTA FIGUERÔA PEIXOTO
Administradora Geral

AGÊNCIA ESTADUAL DE TECNOLOGIA DA INFORMAÇÃO - ATI

Portaria Nº 7/2023. O Diretor-Presidente da Agência Estadual de Tecnologia da Informação - ATI, no uso de suas atribuições legais, e nos termos da Lei nº 7.741 de 23/10/1979, **RESOLVE**:
Art. 1º Designar como Ordenador de Despesa desta Autarquia: I. – ROMERO WANDERLEY GUIMARÃES , Mat. 3143 , Art. 2º Esta Portaria entra em vigor na data de sua publicação retroagindo seus efeitos a 1º de janeiro de 2023.ALLAN RODRIGO DOS SANTOS ARAÚJO, Diretor-Presidente.

FUNDAÇÃO DE APOSENTADORIAS E PENSÕES DOS SERVIDORES DO ESTADO DE PE - FUNAPE

A Diretora-Presidente Interina resolve publicar as Portarias nºs **0529 a 0590** de CONCESSÃO DE PENSÃO POR MORTE, de FEVEREIRO/2023, que se encontram disponíveis, na íntegra, no endereço eletrônico www.funape.pe.gov.br **Ana Maraiza de Sousa Silva** - Diretora-Presidente Interina.

UNIVERSIDADE DE PERNAMBUCO - UPE / REITORIA

A Reitora da Universidade de Pernambuco – UPE assina as seguintes Portarias:
PORTARIA Nº 457/2023 de 13.02.2023
I - Exonerar, a pedido, a servidora DIANDRA DALILA SANTOS SIQUEIRA, mat. nº 16739-4, Assistente Técnica em Gestão Universitária/ Técnica em Informática FDI I A, do Quadro Efetivo de Pessoal desta Universidade, com lotação na UPE Campus Caruaru, a contar de 01.02.2023.
PORTARIA Nº 461/2023 de 13.02.2023
I- Exonerar, a pedido, a servidora MARIA VASCONCELOS CAVALCANTI, mat. nº 12418-6, Médica FDI I C, do Quadro Efetivo de Pessoal desta Universidade, com lotação no Hospital Universitário Oswaldo Cruz - HUOC, a contar de 11.01.2023.
PORTARIA Nº 898/2023 de 16.02.2023
Art. 1º Instaurar Conselho de Sindicância Administrativa para apuração de responsabilidade, no prazo de 20 (vinte) dias, os fatos narrados nos documentos 3044/4984 e 3038402, anexados ao Processo SEI Nº 0340001600303.000003/2023-04, bem como proceder ao exame de outros fatos, ações e omissões, que porventura venham a ser identificados no curso dos trabalhos e que guardem conexão com o objeto presente.
Art. 2º Designar para compor a presente Comissão os servidores: RITA DE CÁSSIA DE MOURA, mat. nº 7453-5, Professora Universitária/ Associada 68H IV D, com lotação no Instituto de Ciências Biológicas - ICB, e MARCELO CASSEB CONTENTINO, mat. nº 14267-8, Professor Universitário/Adjunto MED I A, com lotação na Faculdade de Administração e Direito - FCAP, ambos do Quadro Efetivo de Pessoal desta Universidade para, sob a presidência do primeiro, atuarem na presente apuração.
Art. 3º Determinar que os efeitos desta Portaria sejam a contar de 27.02.2023.
PORTARIA Nº 603/2023 de 16.02.2023
I - Nomear, com a ressalva do Artigo 22 da Lei Complementar Federal nº 101/2000, os concursados descritos abaixo, aprovados no Concurso regido pela Portaria Conjunta SAD/UPE nº 040/2017, de 14.08.2017, Homologado pela Portaria Conjunta SAD/UPE nº 036/2018, de 26.02.2018, e Prorrogado pela Portaria Conjunta SAD/UPE 03/2020, de 20.02.2020.

NOME	CLASSIFICAÇÃO
CAMPUS GARANHUNS	
ANALISTA TÉCNICO EM GESTÃO UNIVERSITÁRIA	
ANALISTA DE SISTEMAS / ÁREA: INFRAESTRUTURA	
KLEDSON DARIO GOMES SILVA	8º
CAMPUS SERRA TALHADA	
ANALISTA TÉCNICO EM GESTÃO UNIVERSITÁRIA	
BIBLIOTECÁRIO	
ANA LIVIA MENDES DE SOUSA	3º
COMPLEXO HOSPITALAR (UPE)	
MÉDICO GINECOLOGISTA E OBSTETRÍCIA	
CAMILA DE ALMEIDA MONTEIRO	49º
ASSISTENTE TÉCNICO EM GESTÃO UNIVERSITÁRIA	
TÉCNICO EM ADMINISTRAÇÃO	
THATIANE DA SILVA ARAUJO	41º
REGIÃO METROPOLITANA DO RECIFE (RMR)	
ANALISTA TÉCNICO EM GESTÃO UNIVERSITÁRIA	
ADMINISTRADOR	
NAYARA BARBOSA RIBEIRO BERNARDO	28º
SECRETARIA EXECUTIVA	
SURAMA KLEIA DA SILVA CHAVES	33º

ASSISTENTE TÉCNICO EM GESTÃO UNIVERSITÁRIA	
ASSISTENTE ADMINISTRATIVO	
ROBERTA MARIA PESSOA MARANHÃO DE LIMA	33º
ANDERSON HENRIQUE DA SILVA (PCD-FÍSICO)	457º

Profª. Drª. **Maria do Socorro de Mendonça Cavalcanti**
REITORA

Licitações e Contratos

ASSEMBLEIA LEGISLATIVA DO ESTADO DE PERNAMBUCO
AVISO DE CHAMAMENTO PÚBLICO
PROCESSO ADMINISTRATIVO N. 001/2023.CPL-ALEPE
A Presidente da Comissão Permanente de Licitações da Assembleia Legislativa do Estado de Pernambuco, torna público para conhecimento dos interessados o **OBJETO**: CREDENCIAMENTO de empresas que prestam serviços de telecomunicações, pelo período de 12 meses, para prestar serviços de telefonia com tecnologia GSM (Global System for Mobile Communications) local (VC1) e longa distância (VC2 e VC3), no sistema ligar pelo-pago, através de plano empresarial, com a disponibilização de estações rebaixa (aparelhos), redes de dados e Internet, de acordo com as normas e regulamentos específicos aplicáveis aos serviços, pelos contratos ou termos de concessão, permissão ou autorização celebrados entre as prestadoras dos serviços e a Agência Nacional de Telecomunicações – ANATEL, contratação de linhas de dados móveis (Modems), com tecnologia mínima 4G, para acesso limitado a Internet, tudo em conformidade com as condições e especificações a seguir apresentadas, para atender às demandas da Assembleia Legislativa do Estado de Pernambuco – ALEPE. O Edital na íntegra pode ser consultado na CPL, em horário de expediente do caso, ou de forma digitalizada através do e-mail **alepe.licita@gmail.com**. Os interessados poderão enviar a documentação de habilitação conforme o disposto neste Edital, de forma física, no endereço: Rua da União, nº 439 – 3º andar, sala 308, Bairro da Boa Vista, Recife, tel. (081) 3183-2521/2448/2383/2106/2447, nos dias úteis, no horário das 08 horas às 16 horas, ou de forma digitalizada através do e-mail **alepe.licita@gmail.com**. Recife, 17 de fevereiro de 2023. Suzana Maria de Aguiar – Presidente da Comissão Permanente de Licitações.

COMPANHIA EDITORA DE PERNAMBUCO - CEPE
PROCESSO LICITATÓRIO Nº 01/2023-PREGÃO ELETRÔNICO 001/2023
AVISO DE LICITAÇÃO: Processo Licitatório 001/2023. Pregão Eletrônico 001/2023. OBJETO: Contratação de empresa especializada para o fornecimento de serviços técnicos especializados no tratamento de conteúdo e informações contemplando a preparação, conversão, separação, indexação, renumeração, inventário de catalogo/documentos em diversos suportes e formatos, manuseio e movimentação de volumes, e fornecimento de solução tecnológica para GED/ECM, contemplando instalação, treinamento e suporte técnico, para atender as necessidades da CEPE, conforme especificações/quantitativos constantes em edital. **RECEBIMENTO DAS PROPOSTAS ATÉ:** 16/03/2023 às 9h. **ABERTURA DAS PROPOSTAS:** 16/03/2023 às 9h15. **INÍCIO DA DISPUTA:** 16/03/2023 às 9h30. O Edital na íntegra poderá ser retirado a partir de 18/02/2023 no site **www.licitacoes-e.com.br**, no site **www.cepe.com.br/licitacoes** ou poderá ser solicitado pelo correio eletrônico **cpl@cepe.com.br**. Recife, 17 de fevereiro de 2023. **Davi Severino de Lima – Pregoeiro**.

DEPARTAMENTO ESTADUAL DE TRÂNSITO DE PERNAMBUCO - DETRAN
EXTRATO DE CONTRATOS, CONVÊNIOS, CREDENCIAMENTOS E TERMOS ADITIVOS
1º TA ao CT de ADESÃO nº 008/2022 à ARP nº 0001.2020. COPLE-UPE.0001.SAD.PE. PARTES: DETRAN/PE e LOCADORA DE VEÍCULOS CAKANDA LTDA. OBJETO: I - Prorrogar vigência; II - Informar dotação orçamentária. VIGÊNCIA: 24/02/2023 a 23/02/2024. VALOR: R$ 411.187,20 (12 meses).

FUND DO PATRIMÔNIO HISTÓRICO E ARTÍSTICO DE PE-FUNDARPE
COMISSÃO PERMANENTE DE LICITAÇÃO – CPL I
RATIFICAÇÃO DE INEXIGIBILIDADE
RECONHEÇO E RATIFICO O **PROCESSO Nº 0171.2023. CPL I.IN.0166.FUNDARPE.** Contratação da BANDA ASAS DA AMÉRICA, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em PEDRAJPE, no dia 18/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: MARCELO BASTOS VALENÇA FILHO, CNPJ: 35.577.066/0001-23. **Valor** R$ 23.000,00. RECONHEÇO E RATIFICO O **PROCESSO Nº 0206.2023.CPL I.IN.0294.FUNDARPE** . Contratação de RENA QUEIROGA, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em OLINDA/PE, no dia 17/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: SOCIEDADE DOS FORROZEIROS PE-DE-SERRA E AI - SOFOPS, CNPJ: **05.584.396/0001-38. Valor** R$ R$ 39.480,00. RECONHEÇO E RATIFICO O **PROCESSO Nº 0219.2023.CPL I.IN.0317.FUNDARPE.** Contratação de CRISTINA AMARAL, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em RECIFE/PE, no dia 18/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: BECK PRODUCOES E EVENTOS LTDA, CNPJ Nº 11.362.371/0001-07. **Valor** R$ 20.800,00. RECONHEÇO E RATIFICO O **PROCESSO Nº 0257.2023.CPL I.IN.0295.FUNDARPE.** Contratação da BANDA BICHO DO MATO, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em SÃO CAETANO/PE, no dia 21/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: CARLOS CAVALCANTI PAES BARRETTO LTDA, CNPJ: 42.583.325/0001-07. **R$ 20.666,00;** RECONHEÇO E RATIFICO O **PROCESSO Nº 0334.2023.CPL I.IN.0332.FUNDARPE.** Contratação de TUCA BARROS, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em GRAVATÁ/PE, no dia 19/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: LAMPEJO ENTRETENIMENTO EIRELI CNPJ 39.537.955/0001-42. R$ 20.000,00; RECONHEÇO E RATIFICO O **PROCESSO Nº 0326.2023.CPL I.IN.0324.FUNDARPE.** Contratação do GRUPO SASSARICO, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em NAZARÉ DA MATA/PE, no dia 18/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: FARIAS EVENTOS E PRODUCOES LTDA, CNPJ: 39.721.242/0001-00. R$ 20.000,00; RECONHEÇO E RATIFICO O **PROCESSO Nº 0325.2023.CPL I.IN.0323.FUNDARPE.** Contratação de NONO GERMANO, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em OLINDA/PE, no dia 18/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: LAMPEJO ENTRETENIMENTO EIRELI CNPJ 39.537.955/0001-42. **R$ 27.000,00;** RECONHEÇO E RATIFICO O **PROCESSO Nº 0327.2023.CPL I.IN.0325.FUNDARPE.** Contratação de MARRON BRASILEIRO, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em GRAVATÁ/PE, no dia 21/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: KNR PRODUÇÕES E EVENTOS LTDA, CNPJ 03.604.473/0001-06. R$ 20.800,00; RECONHEÇO E RATIFICO O **PROCESSO Nº** 0324.2023.CPL I.IN.0322. FUNDARPE. Contratação de ROGÉRIO SOM, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em RIBEIRÃO/PE, no dia 17/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: FRANCISCO S DA COSTA JUNIOR Nº DO CNPJ 32.462.767/0001-90. R$ 34.000,00; RECONHEÇO E RATIFICO O **PROCESSO Nº** 0336.2023.CPL I.IN.0334.FUNDARPE. Contratação da ORQUESTRA MAIA, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em RECIFE/PE, no dia 19/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: EBM CONSTRUÇÃO, ASSESSORIA E SERVIÇOS LTDA CNPJ 33.614.088/0001-76. R$ 20.000,00; RECONHEÇO E RATIFICO O **PROCESSO Nº 0340.2023.CPL I.IN.0346. FUNDARPE.** Contratação de ALLEJANDRO SARNEGADO, para 01 (uma) apresentação durante a programação do CARNAVAL 2023, em OLINDA/PE, no dia 17/02/2023. Fundamentação Legal: Artigo 25, Inciso III, da Lei Federal 8.666/93. Contratado: M CARMO GRAVACOES EVENTOS LTDA Nº DO CNPJ 46.027.877/0001-77. R$ 40.000,00; **RENATA DUARTE BORBA, DIRETORA PRESIDENTE DA FUNDARPE.**

FUND DO PATRIMÔNIO HISTÓRICO E ARTÍSTICO DE PE-FUNDARPE
COMISSÃO PERMANENTE DE LICITAÇÃO – CPL I
ERRATA
Ref. a Publicação no DOE do dia 15/02/2023, Página 10. Ratificação da Inexigibilidade do processo nº **0192.2023.CPL I.IN.0160.FUNDARPE**. Onde se lê: **CARUARU/PE**. Leia-se: **PAULISTA/PE**. Recife, 17/02/2023. **RENATA DUARTE BORBA, DIRETORA PRESIDENTE DA FUNDARPE.**

FUND DO PATRIMÔNIO HISTÓRICO E ARTÍSTICO DE PE-FUNDARPE
COMISSÃO PERMANENTE DE LICITAÇÃO – CPL I
RATIFICAÇÃO DE INEXIGIBILIDADE
RECONHEÇO E RATIFICO O PROCESSO 0280.2023.CPL I.IN.0286.FUNDARPE. Contratação de 07 Maracatus de Baque Solto, para o **3º ENCONTRO ESTADUAL DOS MARACATUS DE BAQUE SOLTO DE PERNAMBUCO**, na programação do CICLO CARNAVALESCO 2023, nos dias 19, 20 e 21 de fevereiro de 2023, no Ilumiara Zumbi na Cidade Tabajara em Olinda/PE e no Terreiro Lenda Alves em Afogados/PE. Fundamentação Legal: Caput do Artigo 25 da Lei Federal 8.666/93. Contratado: **ASSOCIAÇÃO DOS MARACATUS DE BAQUE SOLTO DE PERNAMBUCO**, **CNPJ 12.813.333/0001-19.** Valor: R$ 482.000,00 (quatrocentos e oitenta e dois mil reais). Recife, 16 de fevereiro de 2023. Renata Duarte Borba - Presidente da FUNDARPE.

HOSPITAL AGAMENON MAGALHÃES
EXTRATO DE ATA DE REGISTRO DE PREÇOS
Processo Licitatório CPL/HAM nº 1017.2022 - Pregão Eletrônico nº 0050.2023 Registro de Preços, com validade de 12 (doze) meses, para eventual aquisição de materiais médicos hospitalares - Fica registrado os seguintes itens da empresa vencedora: FR REPRESENTAÇÕES E COMERCIO DE PRODUTOS MEDICOS EIRELI - CNPJ/MF Nº 08.805.088/0001-40, Lotes: 01A, 02A, 01B e 02B), ao valor global de R$ 2.756.080,80 (Dois milhões, setecentos e cinquenta e seis mil reais). As especificações técnicas, bem como os preços unitários dos itens registrados poderão ser observados no ato da homologação de presente processo licitatório.

Jacilene Eustáquio da Silva
Presidente e Pregoeira da CPL

HOSPITAL AGAMENON MAGALHÃES
AVISO DE ADESÃO
O Hospital Agamenon Magalhães comunica a quem interessar possa, que fez adesão a Ata de Registro de Preços, originada do Processo Licitatório nº0018/2022, Pregão Eletrônico nº 0061/2022, promovido pela Comissão Permanente de Licitação do HOSPITAL UNIVERSITÁRIO OSWALDO CRUZ, que tem por objeto a eventual aquisição de material médico hospitalar. Empresa detentora do item:Prima Medical Comercio de Material Médico LTDA, CNPJ:09.340.946/0001-60,(item 01), ao valor total de R$74.030,00(setenta e quatro mil reais).
O Hospital Agamenon Magalhães comunica a quem interessar possa, que fez adesão a Ata de Registro de Preços, originada do Processo Licitatório nº0028/2022, Pregão Eletrônico nº 0063/2022, promovido pela Comissão Permanente de Licitação do HOSPITAL

Figura B.3: Exemplo de Diário Oficial contendo formatação mista de colunas e elementos (Diário Oficial do Estado de Pernambuco).

14 - Ano C • Nº 36 Diário Oficial do Estado de Pernambuco - Poder Executivo Recife, 18 de fevereiro de 2023

UNIVERSITÁRIO OSWALDO CRUZ, que tem por objeto a eventual aquisição de medicamento. Empresa detentora do item:União Química Farmacêutica Nacional S/A. CNPJ:60.665.981/0009-75,(item 04), ao valor total de R$11.700,00(onze mil e setecentos reais).

O Hospital Agamenon Magalhães comunica a quem interessar possa, que foi adesão a Ata de Registro de Preços, originada do Processo Licitatório nº0246/2022, Pregão Eletrônico nº 9046/2022, promovido pela Comissão Permanente de Licitação do HOSPITAL UNIVERSITÁRIO OSWALDO CRUZ, que tem por objeto a eventual aquisição de medicamento. Empresa detentora do item:Farmace Industria Químico-Farmacêutica Cearense LTDA. CNPJ:06.626.333/0001-65,(item 09), ao valor total de R$ 960,00(oito mil cento e sessenta reais).

O Hospital Agamenon Magalhães comunica a quem interessar possa, que foi adesão a Ata de Registro de Preços, originada do Processo Licitatório nº5551/2022, Pregão Eletrônico nº9073/2022, promovido pela Comissão Permanente de Licitação do HOSPITAL UNIVERSITÁRIO OSWALDO CRUZ, que tem por objeto a eventual aquisição de material de almoxarifado. Empresa detentora do item:Nf M Comercio Atacadista LTDA. CNPJ:43.308.383/0001-97,(item 09), ao valor total de R$3.600,00(três mil e seiscentos reais).

O Hospital Agamenon Magalhães comunica a quem interessar possa, que foi adesão a Ata de Registro de Preços, originada do Processo Licitatório nº0816/2022, Pregão Eletrônico nº 9080/2022, promovido pela Comissão Permanente de Licitação do HOSPITAL UNIVERSITÁRIO OSWALDO CRUZ, que tem por objeto a eventual aquisição de material medico hospitalar. Empresa detentora do item:Hpermedic Distribuidora Hospitalar LTDA. CNPJ:28.145.496/0001-90,(item 68), ao valor total de R$4.560,00(quatro mil quinhentos e sessenta reais).

O Hospital Agamenon Magalhães comunica a quem interessar possa, que foi adesão a Ata de Registro de Preços, originada do Processo Licitatório nº0094/2022, Pregão Eletrônico nº 9059/2022, promovido pela Comissão Permanente de Licitação do HOSPITAL UNIVERSITÁRIO OSWALDO CRUZ, que tem por objeto a eventual aquisição de medicamento. Empresa detentora do item:DrogaFonte LTDA. CNPJ:08.779.301/0001-24,(item 17), ao valor total ao R$23.680,00(vinte e três mil oitocentos e oitenta reais).

JACILENE EUSTÁQUIO DA SILVA
PRESIDENTE E PREGOEIRA DA CPL

HOSPITAL AGAMENON MAGALHÃES

ADJUDICAÇÃO E HOMOLOGAÇÃO DE PROCESSO LICITATÓRIO

Processo: Licitatório CPL/HAM nº 0478.2022 - Pregão Eletrônico nº 0023.2022 – Registro de preços, com validade de 12 (doze) meses, para eventual fornecimento de materiais hidráulicos – Empresas Vencedoras: E&M Comércio de Materiais de Construção Eireli, CNPJ/MF nº 24.708.262/0001-73 (itens 03, 07, 09, 12, 14, 15, 17, 23, 26, 28, 30, 34, 36, 38, 39, 44, 45, 47, 48, 51, 52, 53, 57, 59, 62, 63, 64, 66, 71, 72, 76, 82, 85, 92, 93, 95, 97 e 98), ao valor global de R$ 39.560,65 (trinquenta e nove mil, quinhentos e sessenta reais e sessenta e cinco centavos) e Engpc – Engenharia e Assessoria Completas Ltda., CNPJ/MF nº 26.958.042/0001-89 (itens 01, 05, 06, 08, 10, 11, 13, 16, 21, 33, 35, 36, 37, 41, 42, 43, 49, 50, 54, 60, 65, 67, 68, 75, 74, 77, 80, 81, 84, 85, 87, 89, 90, 91 e 96), ao valor global de R$ 99.922,10 (noventa e nove mil, novecentos e vinte e dois reais e dez centavos). Os itens 02, 24, 27, 54, 55, 78, 79 e 86 não foram cotados. Os itens 04, 18, 19, 20, 31, 32, 33, 40, 46, 58, 61, 69, 70, 75, 83, 94, 99 e 100 foram cancelados.

Jacilene Eustáquio da Silva
Presidente e Pregoeira da CPL

HOSPITAL REGIONAL DE LIMOEIRO JOSÉ FERNANDES SALSA

Aviso de Licitação

PROCESSO LICITATÓRIO Nº 2073.2022.CPL-HRL.PE.0086. SES.FES-PE – SEI: 2300211848.000982/2022-39 Objeto: Aquisição de **MATERAIS DE LIMPEZA (LAVANDERIA)**. Valor máximo aceitável **R$ 75.335,4494 (setenta e cinco mil trezentos e trinta e cinco reais, quatro mil quatrocentos e quatro décimos de milésimos)**. Entrega de Propostas até 10.03.2023 às 09h00min. Início dos Lances: 10.03.2023 às 09h30min. Horário de Brasília.

PROCESSO LICITATÓRIO Nº 3003.2022.CPL.HRL.PE.0086. SES.FES-PE – SEI: 2300211848.000980/2022-15 Objeto: **CONTRATAÇÃO DE EMPRESA ESPECIALIZADA EM SERVIÇOS DE MANUTENÇÃO PREVENTIVA E CORRETIVA EM APARELHO DE RAIO X E PROCESSADORA DE IMAGEM PARA RAIO X (COM REPOSIÇÃO DE PEÇAS)**. Valor máximo aceitável **R$ 304.960,5054 (trezentos e quatro mil novecentos e sessenta e cito reais, cinco mil seiscentos e quatro décimos de milésimos)**. Entrega de Propostas até 09.03.2023 às 09h00min. Início dos Lances: 09.03.2023 às 09h30min. Horário de Brasília. Os editais, na íntegra, poderão ser retirados na página eletrônica: www.pelotagrado.pe.gov.br. Outras informações pelo e-mail: licitacaohrl2013@outlook.com. ou, pelo Fone: (081) 3628 3805/8800. Limoeiro, 17.02.2023. Roberta Maria de Santana – Pregoeira.

INSTITUTO AGRONÔMICO DE PERNAMBUCO – IPA

1º TA IPA/MU Nº 25/2022. CT Nº 1/2022 celebrado com **DIBASA COMERCIO E SERVIÇOS TECNICOS LTDA.** CNPJ: 11.836.848/0001-71; Objeto: prorrogação do prazo de vigência do contrato supracitado, por mais 12 (doze) meses, pelo período de 27/01/2023 a 26/01/2024. Assinatura/Início: 30/12/2022. **BARTOLOMEU MONTEIRO – Diretor Presidente.**

INSTITUTO DE RECURSOS HUMANOS DE PERNAMBUCO - IRH

EXTRATOS DE ATA DE REGISTRO DE PREÇOS

PI.Nº0122.2022.CPL-E validade: 12 meses; obj: **MEDICAMENTOS DE SUPORTE** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **CHRISPIM NEDI CARRILHO EIRELI** CNPJ: **31600214/70-9**; ITEM 5; Valor total: **R$ 13.742,46**. Recife, 09/02/2023.

PI.Nº0122.2022.CPL-E validade: 12 meses; obj: **MEDICAMENTOS DE SUPORTE** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **UNI HOSPITALAR LTDA.** CNPJ: **07.484.373/0001-24**; ITEM 1; Valor total: **R$: 189.790,0000**. Recife, 09/02/2023.

PI.Nº0122.2022.CPL-E validade: 12 meses; obj: **MEDICAMENTOS DE SUPORTE** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **TECHPHARMA HOSPITALAR COMERCIO, IMPORTAÇÃO E EXPORTAÇÃO EIRELI**. CNPJ: **35.267.853/0001-25**; ITEM 3; Valor total: **R$: 17.999,95**. Recife, 08/02/2023.

PI.Nº0122.2022.CPL-E validade: 12 meses; obj: **MEDICAMENTOS DE SUPORTE** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **UNIKA DISTRIBUIDORA DE MEDICAMENTOS LTDA**. CNPJ: **35.253.360/0001-60**; ITEM 1; Valor total: **R$: 26.275,9200**. Recife, 09/02/2023.

PI.Nº0155.2022.CPL-E validade: 12 meses; obj: **DIETAS ENTERAIS EM SISTEMA FECHADO** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **ARSERVE PHARMA EPP LTDA.** CNPJ: **43.019.101/0001-70**; ITEM 8,20; Valor total: **R$: 36.482,00**. Recife, 14/02/2023.

PI.Nº0165.2022.CPL-E validade: 12 meses; obj: **DIETAS ENTERAIS EM SISTEMA FECHADO** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **CENUT DISTRIBUIDORA DE PRODUTOS ALIMENTICIOS DE SAUDE LTDA.** CNPJ: **30.595.447/0002-36**; ITEM 1; Valor total: **R$: 46.895,80**. Recife, 14/02/2023.

PI.Nº0185.2022.CPL-E validade: 12 meses; obj: **DIETAS ENTERAIS EM SISTEMA FECHADO** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **CL COMERCIO DE MATERIAIS MÉDICOS HOSPITALARES LTDA-EPP**. CNPJ: **13.441.551/0002-01**; ITEM 14,15,16,17,18,21,22,24; Valor total: **R$: 146.824,73**. Recife, 09/02/2023.

PI.Nº0185.2022.CPL-E validade: 12 meses; obj: **DIETAS ENTERAIS EM SISTEMA FECHADO** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **NUTRI HOSPITALAR LTDA.** CNPJ: **10.782.960/0001-70**; ITEM 10; Valor total: **R$: 30.600,00**. Recife, 08/02/2023.

PI.Nº0185.2022.CPL-E validade: 12 meses; obj: **DIETAS ENTERAIS EM SISTEMA FECHADO** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **VITALE COMERCIO S.A.** CNPJ: **07.160.018/0002-28**; ITEM: 3,5,7,9,11; Valor total: **R$: 266.884,4760**. Recife, 08/02/2023.

PI.Nº0185.2022.CPL-E validade: 12 meses; obj: **MATERIAL CIRURGICO UROLÓGICO E UROGINECOLÓGICO** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **CIENLABOR INDUSTRIA E COMERCIO LTDA.** CNPJ: **02.014.280/0002-08**; ITEM 1,2; Valor total: **R$: 162.000,00**. Recife, 14/02/2023.

PI.Nº0185.2022.CPL-E validade: 12 meses; obj: **MATERIAL CIRURGICO UROLÓGICO E UROGINECOLÓGICO** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **MEDICAL NORDESTE IMPORTACAO E DISTRIBUICAO DE PRODUTOS MEDICOS LTDA.** CNPJ: **17.817.514/0001-55**; **LOTE 1**: ITEM 1; LOTE 2: ITEM 1,2,3,4,5,6,7,8,9,10,11,12,13,14, 15,16; LOTE 4º ITEM 1,2,3; LOTE 4B ITEM 1,2,3 Valor total: **R$: 2.627.744,10**. Recife, 14/02/2023.

PI.Nº0122.2022.CPL-E validade: 12 meses; obj: **MEDICAMENTOS ONCOLÓGICOS** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **UNIKA DISTRIBUIDORA DE MEDICAMENTOS LTDA.** CNPJ: **35.253.360/0001-60**; ITEM 1; Valor total: **R$: 26.275,9200**. Recife, 09/02/2023.

PI.Nº0181.2022.CPL-E validade: 12 meses; obj: **MATERIAL MÉDICO HOSPITALAR** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **CARLA DE OLIVEIRA CORREA.** CNPJ: **08.583.329/0001-08**; ITEM 10; Valor total: **R$: 24.360,00**. Recife, 16/02/2023.

PI.Nº0185.2022.CPL-E validade: 12 meses; obj: **MATERIAL MÉDICO HOSPITALAR** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **CRUZSUL COMERCIAL LTDA.** CNPJ: **10.877.170/0001-43**; ITEM 15,16,25,26; Valor total: **R$: 75.019,50**. Recife, 16/02/2023.

PI.Nº0181.2022.CPL-E validade: 12 meses; obj: **MATERIAL MÉDICO HOSPITALAR** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **ENDOSURGICAL COMERCIO, REPRESENTAÇÃO, IMPORTAÇÃO E EXPORTAÇÃO DE MATERIAIS E EQUIPAMENTOS MEDICOS E ODONTOLOGICOS LTDA.** CNPJ: **08.713.823/0001-55**; ITEM 10,13; Valor total: **R$: 412.230,00**. Recife, 16/02/2023.

PI.Nº0181.2022.CPL-E validade: 12 meses; obj: **MATERIAL MÉDICO HOSPITALAR** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **JRV HOSPITALAR COMERCIO E REPRESENTACAO EIRELI**. CNPJ: **40.829.708/0001-74**; ITEM 24; Valor total: **R$: 2.160,00**. Recife, 16/02/2023.

PI.Nº0181.2022.CPL-E validade: 12 meses; obj: **MATERIAL MÉDICO HOSPITALAR** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **MJB COMERCIO DE MATERIAIS MEDICO HOSPITALARES LTDA.** CNPJ: **08.014.554/0001-00**; ITEM 2,4,5,6,7,8,9,11,14,20,21; Valor total: **R$: 347.571,22**. Recife, 15/02/2023.

PI.Nº0181.2022.CPL-E validade: 12 meses; obj: **MATERIAL MÉDICO HOSPITALAR** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. Ata nº **26776235 /2022 – IRH**, Fornec: **PONTO 71 IMPORTAÇÃO E EXPORTAÇÃO EIRELI**. CNPJ: **17.035.479/0001-02**; ITEM 17,19; Valor total: **R$: 58.025,00**. Recife, 16/02/2023.

Recife, 16 de FEVEREIRO de 2023
JOÃO VICTOR FALCÃO DE ANDRADE
Diretor Presidente

INSTITUTO DE RECURSOS HUMANOS DE PERNAMBUCO - IRH

EXTRATOS DOS ADITIVOS DE TERMOS DE CREDENCIAMENTOS

6º Termo Aditivo ao Termo de Credenciamento Nº **036/2017** Contratada: **HOSPITAL MEMORIAL ARCOVERDE**. CNPJ: **10.237.227/0001-30**; objeto: prorrogação da prestação de Serviços de Assistência à Saúde; Nova vigência: 01/11/2022 a 31/10/2023. Recife, 01/11/2022

3º Termo Aditivo ao Termo de Credenciamento Nº **022/2018** Contratada: **CEM - CLINICA DE ESPECIALIDADES MÉDICAS LTDA.** CNPJ: **09.237.383/0001-03**; objeto: prorrogação da prestação de Serviços de Assistência à Saúde; Nova vigência: 01/08/2022 a 31/07/2023. Recife, 01/08/2022

Recife, 30 de DEZEMBRO de 2022
RICARDA SAMARA DA SILVA BEZERRA
Diretora Presidente

EXTRATOS DE TERMOS DE CREDENCIAMENTOS

Termo de Credenciamento nº **028/2021**, Contratada: **OFTALMAX HOSPITAL DE OLHOS LTDA.** CNPJ sob o nº **02.166.836/0002-77**; Objeto: **Serviços de Assistência à Saúde**; Vigência: 12 meses a partir de 25/11/2021; Valor Total: R$ **4.280.000,00**. Recife, 10/02/2022

Termo de Credenciamento nº **001/2022**, Contratada: **B&M OFTALMOLOGIA LTDA.** CNPJ sob o nº **12.943.564/0001-11**; Objeto: **Serviços de Assistência à Saúde**; Vigência: 12 meses a partir de 10/02/2022; Valor Total: R$ **864.000,00**. Recife, 10/02/2022

Recife, 30 de DEZEMBRO de 2022
RICARDA SAMARA DA SILVA BEZERRA
Diretora Presidente

EXTRATOS DE TERMOS DE CONTRATO DE DISPENSA

Nº **160**, validade: 180 dias; obj: **MEDICAMENTOS DIVERSOS** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE. **PL: 197.2022; DISPENSA: 20.2022**, Fornec: **EXOMED COMERCIO ATACADISTA DE MEDICAMENTOS LTDA.** CNPJ: **12.602.833/0001-94**; ITEM 19, 62, 66; Valor total: **R$: 37.918,74**. Recife, 16/12/2022.

Nº **98**, validade: 180 dias; obj: **MEDICAMENTOS DIVERSOS** visando atender à demanda do Hospital dos Servidores do Estado de Pernambuco – HSE, **PL: 0157.2022; DISPENSA: 20.2022** Fornec: **ONCOEXO DISTRIBUIDORA DE MEDICAMENTOS LTDA.** CNPJ: **00.502.023/0001-06**; ITEM 38,52; Valor total: **R$: 65.129,08**. Recife, 16/12/2022.

Recife, 30 de DEZEMBRO de 2022
RICARDA SAMARA DA SILVA BEZERRA
Diretora Presidente

ERRATA AO TERMO ADITIVO DO TERMO DO CREDENCIAMENTO: **DIAGNÓSTICO DA AMÉRICA SA CERPE**, CNPJ **61.486.650/0886-12**. **Onde se lê**: " DIAGNÓTICO DA AMÉRICA SA CERPE, inscrita no CNPJ (MF) sob o nº 61.486.650/884-12." **Leia-se**: " DIAGNOSTICO DA AMERICA SA CERPE, inscrita no CNPJ (MF) sob o nº 61.486.650/0884-12."

ERRATA DA PUBLICAÇÃO 20/08/2022: **DOUTOR ATENDE GESTÃO E ADMINISTRAÇÃO DE CONSULTÓRIOS LTDA.** CNPJ: **12.024.227/0005-93**. **Onde se lê**: " O prazo de vigência do presente instrumento será de 01/01/2022 a 31/03/2023 podendo ser prorrogado por interesse das partes". **Leia-se**: " O prazo de vigência do presente instrumento será de 30/12/2022 até 29/03/2023 podendo ser prorrogado por interesse das partes."

Recife, 30 de DEZEMBRO de 2022
RICARDA SAMARA DA SILVA BEZERRA
Diretora Presidente

INSTITUTO DE RECURSOS HUMANOS DE PERNAMBUCO - IRH

EXTRATOS DE ADESÃO A ATA DE REGISTRO DE PREÇOS

PI.Nº**3218/2007/2022**; validade: 12 meses; obj: **materiais hospitalares**, para Atender as necessidades do Hospital dos Servidores do Estado de Pernambuco – HSE, Termo de Adesão nº **26776761/2022** Fornec: **COMERCIAL CIRURGICA RIOCLARENSE LTDA.** CNPJ: **47.729.178/0002-20**; ITEM 1,2; Valor total: **R$ 41.120,0000**. Recife, 17/02/2023.

PI.Nº**3218/1466/2022**; validade: 12 meses; obj: **materiais hospitalares**, para Atender as necessidades do Hospital dos Servidores do Estado de Pernambuco – HSE, Termo de Adesão nº **26776761/2022** Fornec: **CALL MED COMERCIO DE MEDICAMENTOS E REPRESENTAÇÃO LTDA - ME.** CNPJ: **05.106.015/0001-62**; ITEM 9; Valor total: **R$ 71.040,0000**. Recife, 17/02/2023.

PI.Nº**3132/0507/2022**; validade: 12 meses; obj: **MEDICAMENTOS MANIPULADOS** para Atender as necessidades do Hospital dos Servidores do Estado de Pernambuco - HSE, Termo de Adesão nº **26706761/2022** Fornec: **DERMATOFLORA LTDA.** CNPJ: **17.018.735/0001-07**; ITEM 1; Valor total: **R$ 1.350,00**. Recife, 17/02/2023.

PI.Nº**3132/0507/2022**; validade: 12 meses; obj: **MEDICAMENTOS MANIPULADOS** para Atender as necessidades do Hospital dos Servidores do Estado de Pernambuco - HSE, Termo de Adesão nº **26706761/2022** Fornec: **DERMATOFLORA LTDA.** CNPJ: **17.018.735/0001-07**; ITEM 1; Valor total: **R$ 1.350,00**. Recife, 17/02/2023.

Recife, 17 de FEVEREIRO de 2023
DJOÃO VICTOR FALCÃO DE ANDRADE
Diretor Presidente

LABORATÓRIO FARMACÊUTICO DO ESTADO DE PERNAMBUCO GOVERNADOR MIGUEL ARRAES S/A - LAFEPE

COMISSÃO PERMANENTE DE LICITAÇÃO-CPL

AVISO DE ADJUDICAÇÃO

SEI Nº 0066407867.000933/2022-48

ADJUDICO o objeto do **PROCESSO LICITATÓRIO Nº 091/2022 – PREGÃO ELETRÔNICO Nº 048/2022**, em favor das empresas: **VTA MACHADO DE ARRUDA E CIA LTDA** (CNPJ: 16.067.433/0001-30), **LOTE 1-R$ 64.490,00** (Sessenta e quatro mil e quatrocentos reais), **LOTE 2-R$ 16.160,00** (Dezesseis mil e cem reais), **LOTE 3-R$ 31.315,00** (Trinta e um mil, trezentos e quinze reais), **LOTE 5-R$ 27.875,00** (Vinte e sete mil, oitocentos e setenta e cinco reais), **COMODORO COMERCIAL E NUTRIÇÃO LTDA.** ME. (CNPJ: 10.461.277/0001-75), **LOTE 4-R$ 47.500,00** (Quarenta e sete mil e quinhentos reais). Recife, 17/02/2023. Roziane Brito – Pregoeira.

RATIFICAÇÃO DE INEXIGIBILIDADE

PROCESSO Nº 0065/2022 – INEXIGIBILIDADE Nº 011/2023
SEI Nº 0066407879.000247/2022-14

Reconheço e ratifico com base na Lei Federal nº 13.303/16, para fins do disposto no art. 30, caput, da lei supra, e à vista da justificativa do setor demandante e Parecer Jurídico (Doc. SEI 33462895), a inexigibilidade de licitação em epígrafe, cujo objeto é a **Prestação de serviço com fornecimento de peças** para o equipamento **CROMATÓGRAFO GASOSO-CG** modelo **8860 GC SYSTEM** da empresa **AGILENT TECHNOLOGIES BRASIL LTDA**, através da contratação direta da empresa **AGILENT TECHNOLOGIES BRASIL LTDA** (CNPJ nº 03.260.230/0006-08), o valor total da contratação será de R$ 51.404,17 (Cinquenta e um mil, quatrocentos e quatro reais e dezessete centavos). Recife, 17/02/2023. Belq Sarma – Diretora Técnica Industrial.

POLÍCIA MILITAR DE PERNAMBUCO

DCC/DEAJA – Terno de Contrato

TC nº 005/2023. Proc. 021.2022.CPLPE. 0057/PMPE serviços de montagem e desmontagem plataformas. Empresa: Polygon Locação. 39.647.536/0001-91. Vigência: 09/02/23 a 08/02/24. Valor R$ 75.000,00.

POLÍCIA MILITAR DE PERNAMBUCO

Aviso de Licitação

Processo Licitatório Nº 0005.2023.CPLPE.0002.PMPE-CPL/ Capital - Registro de Preços para o Fornecimento eventual de Gêneros Alimentícios visando atender às necessidades das Unidades da PMPE (CREED, BPGD, BPCHOQUE, RPMON, e ACG) - valor: R$ 1.553.803,1621 - Recebimento das Propostas: até 06/MAR/2023 às 10h00 (Horário de Brasília). Disputa de Preços: 06/MAR/2023 às 10h30min OBS: O edital na íntegra poderá ser retirado na CPL/Capital, sito a Rua Antonio Bezerra s/nº, Derby – Recife/PE, nos sites www.pmtegrado.pe.gov.br e www.pm.pe.gov.br ou pelo e-mail cpl@pm.pe.gov.br. Informações: Fones: (81) 3181 1126/1203. Recife, 17/FEV/2023 – André Felipe Araújo P. do Nascimento – Ten Cel PM – Presidente da CPL/ Capital.

PROCURADORIA GERAL DO ESTADO

2ºTA.CT.05/20-RG-01/23-ORACLE DO BRASIL SISTEMAS LTDA – prorr. por 12 meses. 01/01/23 a 31/12/24. 23NE31 R$ 2.479,68 e 23NE32 R$12,96. 1ºTA.CT.09/22-RG-02/23-J.R. COM. SERV. CLIMATIZAÇÃO – prorr. por 12 meses. 09/02/23 a 07/02/24. 23NE62 R$9.084,80. 23NE100 R$276,74 e 23NE101 e R$420,78. 2ºTA.CT.04/21-RG-03/23-MBM SEGURADORA S/A – prorr. por 12 meses. 05/02/23 a 04/02/24. 23NE28 R$995,97. 1ºTA.CT.10/22-RG-04/23-J.BATISTA S. FERREIRA – prorr. por 12 meses. 16/02/23 a 15/02/24. 23NE77 R$17.585,60.

SECRETARIA DE ADMINISTRAÇÃO

SECRETARIA DE ADMINISTRAÇÃO - ÓRGÃO CENTRAL DE PATRIMÔNIO DO ESTADO

AUTORIZAÇÃO DE USO DE IMÓVEL ESTADUAL - SEI 000120012/2022/23/2023-48

Autorizante: Secretaria de Administração. Autorizatário: Fundação da Cultura da Cidade do Recife - FCCR. Objetos: Avenida Agamenon Magalhães, nº 5091, Santo Amaro, Recife/PE e Avenida Agamenon Magalhães, s/n, Salgadinho, Olinda/PE (área de estacionamento). Vigência: de 10 a 12 de fevereiro de 2023. Finalidade: área de apoio e circulação da equipe de trabalho, montagem de estrutura, operação de trânsito e estacionamento para evento cultural.

SECRETARIA DE ADMINISTRAÇÃO

AVISO DE ADIAMENTO DE LICITAÇÃO

PROCESSO Nº 0002.2023.COPLE LPE.0002.SAD.SEDUC
PROCESSO SEI Nº 1400005288.000329/2022-71

Objeto: Formação de Registro de Preços para a contratação de prestação de serviços de fornecimento de alimentação escolar (in loco e transportada), lanches e almoços, com aquisição e aprovisionamento de todos os gêneros alimentícios e demais insumos, para os estudantes de 2/3 (duzentas e oitenta) unidades de ensino, nos termos da legislação vigente. Comunica-se o adiamento Sine Die da sessão de abertura agendada para o dia 24/02/2023, em razão da solicitação emanada pela SEE no Despacho 08, Doc. SEI nº 33574777, fundamentada no Ofício nº 1477/2023 – TCE-PE/GOVI assinado pelo Tribunal de Contas de Pernambuco e recebido pela COPLE I em 16/02/2023. Recife, 17/02/2023. Mário Borges. Pregoeiro/Presidente da COPLE I.

SECRETARIA DE ADMINISTRAÇÃO

EXTRATO DE CONTRATO

Processo Nº 0156.2021.COPLE-I/PE.0135.SAD. Modalidade/ Pregão Eletrônico Nº 0125.2021. Objeto Nat.: Outros serviços. Objeto Descr.: Prestação de serviços de copeiragem para a SAD. Contrato nº 006/2023. Contratada: **SERCOSERV SERVIÇOS TERCEIRIZADOS EIRELI; CNPJ nº 08.717.223/0001-06**. Valor Contratado: R$ **80.905,00**. Recife, 17 de fevereiro de 2023.

SECRETARIA DE DEFESA SOCIAL

EXTRATO DO CONTRATO Nº 173/2022-GAB/SDS que entre si celebram o Estado de Pernambuco, através da Secretaria de Defesa Social CNPJ nº 02.960.840/0001-06 e VECTRA WORK INDUSTRIA E COMERCIO DE UNIFORMES E EQUIPAMENTOS DE PROTEÇÃO INDIVIDUAL LTDA, CNPJ nº 11.694.766/0001-44. Objeto: Aquisição de capacetes para o combate a distúrbio urbano e fluvial, destinado ao Corpo de Bombeiros Militar de Pernambuco. Utilização da ATA DE REGISTRO DE PREÇOS Nº 069/2022-SLC, decorrente do PREGÃO ELETRÔNICO Nº 002/12021 – CPL II, PROCESSO LICITATÓRIO Nº 003/12021 – CPL II. Fundamento Legal: Lei 10.520/2002, Decretos Estaduais 32.536/2009 e 42.530/2015, aplicando-se subsidiariamente a Lei 8.666/1993. Valor Total: R$ 425.100,00 (quatrocentos

Figura B.4: Exemplo de Diário Oficial onde, em contraste com a Figura B.3, é observada divergência de formatação dentro de um mesmo Diário Oficial (Diário Oficial do Estado de Pernambuco).

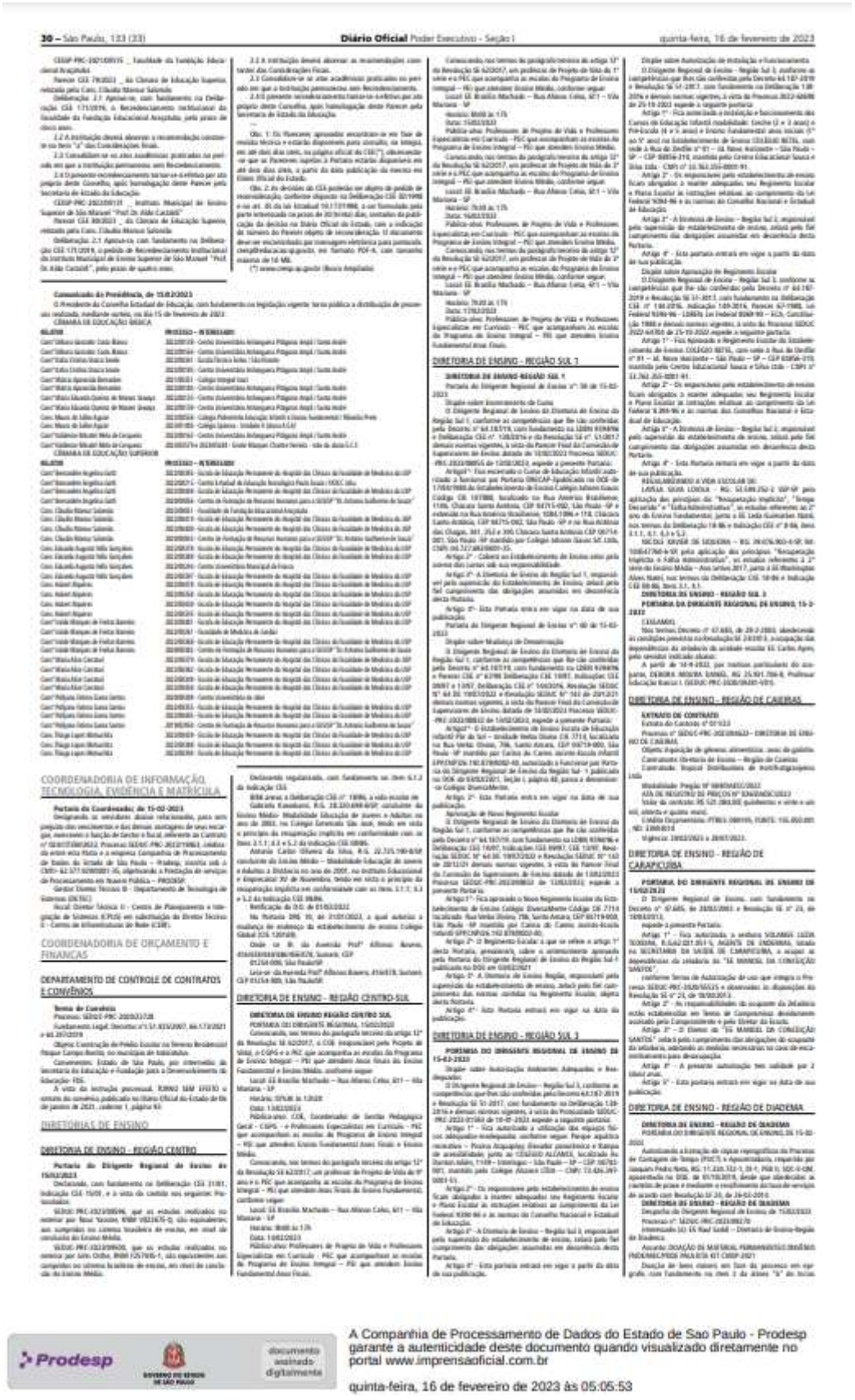


Figura B.5: Exemplo de Diário Oficial contendo formatação mista de colunas e elementos (Diário Oficial do Estado de São Paulo).

98 Maceió - quinta-feira 16 de fevereiro de 2023 Edição Eletrônica Certificada Digitalmente conforme LEI Nº 7.397/2012 Diário Oficial Estado de Alagoas

CPF nº 108.630.194-39, para exercer o cargo, de provimento em comissão, de Assessor Técnico, Nível: AST-3, da Secretaria do Trabalho e Emprego, do Serviço Civil do Poder Executivo, criado pela Lei Delegada nº 48, de 30 de dezembro de 2022.

PALÁCIO REPÚBLICA DOS PALMARES, em Maceió, 15 de fevereiro de 2023, 207º da Emancipação Política e 135º da República.

PAULO SURUAGY DO AMARAL DANTAS
Governador

DECRETO Nº 89.041, DE 15 DE FEVEREIRO DE 2023.

O GOVERNADOR DO ESTADO DE ALAGOAS, no uso de suas atribuições, RESOLVE exonerar, a pedido, EDER RAFAEL DE QUEIROZ BARROS, CPF nº 078.136.594-57, do cargo, de provimento em comissão, de Assessor Técnico de Atendimento, Nível AST-4, da Agência de Modernização da Gestão de Processos - AMGESP, do Serviço Civil do Poder Executivo.

PALÁCIO REPÚBLICA DOS PALMARES, em Maceió, 15 de fevereiro de 2023, 207º da Emancipação Política e 135º da República.

PAULO SURUAGY DO AMARAL DANTAS
Governador

DECRETO Nº 89.042, DE 15 DE FEVEREIRO DE 2023.

O GOVERNADOR DO ESTADO DE ALAGOAS, no uso das atribuições que lhe confere o inciso XIV do art. 107 da Constituição Estadual, RESOLVE nomear JOSÉ DE MACEDO FERREIRA, CPF nº 049.533.714-53, para exercer o cargo, de provimento em comissão, de Secretário Executivo de Articulação Política – Governadoria do Agreste, Nível SEE do Gabinete Civil, do Serviço Civil do Poder Executivo, na forma da Lei Delegada nº 50, de 6 de fevereiro de 2023.

PALÁCIO REPÚBLICA DOS PALMARES, em Maceió, 15 de fevereiro de 2023, 207º da Emancipação Política e 135º da República.

PAULO SURUAGY DO AMARAL DANTAS
Governador

DECRETO Nº 89.043, DE 15 DE FEVEREIRO DE 2023.

O GOVERNADOR DO ESTADO DE ALAGOAS, no uso das atribuições que lhe confere o inciso XIV do art. 107 da Constituição Estadual, RESOLVE nomear ELI MARIO MAGALHÃES MORAES, CPF nº 111.239.824-49, para exercer o cargo, de provimento em comissão, de Secretário Executivo de Articulação Social – Governadoria do Agreste, Nível SEE, do Gabinete Civil, do Serviço Civil do Poder Executivo, na forma da Lei Delegada nº 50, de 6 de fevereiro de 2023.

PALÁCIO REPÚBLICA DOS PALMARES, em Maceió, 15 de fevereiro de 2023, 207º da Emancipação Política e 135º da República.

PAULO SURUAGY DO AMARAL DANTAS
Governador

DECRETO Nº 89.044, DE 15 DE FEVEREIRO DE 2023.

O GOVERNADOR DO ESTADO DE ALAGOAS, no uso de suas atribuições, considerando o disposto no inciso XIV do art. 107 da Constituição Estadual, RESOLVE nomear JULLYENE CRISTINE SANTOS LINS, CPF nº 995.182.014-04, para exercer o cargo, de provimento em comissão, de Assessor Especial da SEAGRI, Nível ASEG, da Secretaria de Estado de Agricultura e Pecuária – SEAGRI, do Serviço Civil do Poder Executivo, na forma da Lei Delegada nº 52, de 10 de fevereiro de 2023.

PALÁCIO REPÚBLICA DOS PALMARES, em Maceió, 15 de fevereiro de 2023, 207º da Emancipação Política e 135º da República.

PAULO SURUAGY DO AMARAL DANTAS
Governador

DECRETO Nº 89.045, DE 15 DE FEVEREIRO DE 2023.

O GOVERNADOR DO ESTADO DE ALAGOAS, no uso de suas atribuições, considerando o que estabelece o art. 96, da Lei nº 5.247, de 26 de julho de 1991, com a redação que lhe foi dada pela Lei nº 5.700, de 16 de junho de 1995, e tendo em vista o que consta no Processo Administrativo nº E:41506-0000000604/2022, RESOLVE ceder o servidor DELPHINO DE OLIVEIRA CAVALCANTE, CPF nº 347.048.684-00, ocupante do cargo de Agente Administrativo, matrícula nº 843-5, lotado na Secretaria de Estado do Planejamento, Gestão e Patrimônio - SEPLAG, ao Instituto de Tecnologia em Informática e Informação do Estado de Alagoas – ITEC, sem ônus para o órgão de origem, até o término do atual período administrativo governamental.

PALÁCIO REPÚBLICA DOS PALMARES, em Maceió, 15 de fevereiro de 2023, 207º da Emancipação Política e 135º da República.

PAULO SURUAGY DO AMARAL DANTAS
Governador

DECRETO Nº 89.046, DE 15 DE FEVEREIRO DE 2023.

O GOVERNADOR DO ESTADO DE ALAGOAS, no uso de suas atribuições, considerando o que estabelece o art. 96, da Lei nº 5.247, de 26 de julho de 1991, com a redação que lhe foi dada pela Lei nº 5.700, de 16 de junho de 1995, e tendo em vista o que consta no Processo Administrativo nº E:41506-0000000604/2022, RESOLVE ceder o servidor EDNOR VIEIRA DE LIMA, CPF nº 163.898.754-87, ocupante do cargo de Assistente de Administração, matrícula nº 43.143-5, lotado na Secretaria de Estado do Planejamento, Gestão e Patrimônio - SEPLAG, ao Instituto de Tecnologia em Informática e Informação do Estado de Alagoas – ITEC, sem ônus para o órgão de origem, até o término do atual período administrativo governamental.

PALÁCIO REPÚBLICA DOS PALMARES, em Maceió, 15 de fevereiro de 2023, 207º da Emancipação Política e 135º da República.

PAULO SURUAGY DO AMARAL DANTAS
Governador

DECRETO Nº 89.047, DE 15 DE FEVEREIRO DE 2023.

O GOVERNADOR DO ESTADO DE ALAGOAS, no uso da atribuição que lhe confere o inciso XVI do art. 107 da Constituição Estadual, tendo em vista o contido no Parecer AL PREVIDÊNCIA SUBPGE 16509431 e no Despacho PGE COOPA 16547868, aprovado pelo Despacho PGE GPG 16581355, todos da Procuradoria Geral do Estado, e o que mais consta do Processo Administrativo nº E:02000.0000000138/2019,

Figura B.6: Exemplo de Diário Oficial contendo formatação de coluna dupla, porém distinta do observado nas Figuras B.1 e B.2 (Diário Oficial do Estado de Alagoas).

DIÁRIO OFICIAL DA UNIÃO - Seção 1 ISSN 1677-7042 Nº 35, sexta-feira, 17 de fevereiro de 2023

SOLUÇÃO DE CONSULTA Nº 98.019, DE 30 DE JANEIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM: 8426.49.90
Mercadoria: Máquina autopropulsada sobre rodas com pneus maciços de borracha, de cabine única, utilizada para içamento e movimentação em curtas distâncias de sucata, toras e outros tipos de materiais, por meio de garras ou outras ferramentas fixadas na ponta do braço de trabalho articulado (lança); com capacidade de elevação máxima de 4,8 t; não apropriada para o transporte de cargas; cujos mecanismos de propulsão, de manipulação de materiais e chassi formam um corpo único homogêneo.
Dispositivos Legais: RGI 1, RGI 6 e RGC 1 da NCM constante da TEC, aprovada pela Res. Gecex nº 272, de 2021, e da Tipi, aprovada pelo Dec. nº 11.158, de 2022; e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e pelas IN RFB nº 1.788, de 2018, nº 2.052, de 2021, e alterações posteriores.

MARCO ANTÔNIO RODRIGUES CASADO
Presidente da 5ª Turma

SOLUÇÃO DE CONSULTA Nº 98.020, DE 30 DE JANEIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM: 2710.12.90
Ex Tipi: sem enquadramento.
Mercadoria: Mistura líquida de números de cadeia, constituída aproximadamente por 60% de n-pentano e 40% de isopentano (hidrocarbonetos leves de cadeia aberta, saturada, com baixo ponto de ebulição), utilizada como agente expansor na formulação de espumas poliméricas, acondicionada em isotanque de 14.000 kg.
Dispositivos Legais: RGI 1, RGI 6 (Nota de subposição 4 do Cap. 27) e RGC 1 da NCM constante da TEC, aprovada pela Res. Gecex nº 272, de 2021, e da Tipi, aprovada pelo Dec. nº 11.158, de 2022; e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e pelas IN RFB nº 1.788, de 2018, nº 2.052, de 2021, e alterações posteriores.

MARCO ANTÔNIO RODRIGUES CASADO
Presidente da 5ª Turma

SOLUÇÃO DE CONSULTA Nº 98.021, DE 30 DE JANEIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM 9031.80.99
Mercadoria: Aparelho para identificação de falhas ou defeitos em peças por meio de detecção de microvazamentos, através da medição de fuga por microprensão diferencial, utilizando circuito pneumático otimizado e conversor de pressão diferencial (Pa) em unidade de vazamento (cm³/min), com display com tela sensível ao toque, entradas para comunicação em redes de dados, entrada para o fluido de teste pressurizado e saídas por onde o fluido pressurizado, devidamente ajustado aos parâmetros de teste, é disponibilizado para ser conduzido por dispositivos auxiliares (não apresentados com o aparelho) onde a peça a ser testada é colocada. É utilizado principalmente para detecção de falhas em peças automotivas, manufaturadas, linha branca, metais sanitários, embalagens, produtos farmacêuticos e equipamentos médicos.
Dispositivos Legais: RGI 1, RGI 6 e RGC 1 da NCM constante da TEC, aprovada pela Res. Gecex nº 272, de 2021, e da Tipi, aprovada pelo Dec. nº 11.158, de 2022; e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e pelas IN RFB nº 1.788, de 2018 e nº 2.052, de 2021, e alterações posteriores.

MARCO ANTÔNIO RODRIGUES CASADO
Presidente da 5ª Turma

SOLUÇÃO DE CONSULTA Nº 98.022, DE 30 DE JANEIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM 9026.80.00
Mercadoria: Fluxômetro digital para medição de vazão mássica de gases, utilizando tecnologia CMOSens, incorporando unidades de estanqueidade e medidor de volume, com algoritmo de cálculo de média corrente, medição dinâmica do fluxo, apto a trabalhar em linhas pressurizadas de até 1 MPa, com interface touchscreen, , comunicação serial Ethernet (RJ45) - TELNET e HTTPS e RS-232-MODBUS, com interface de saída analógica 4~20 mA e 1~5 Vdc, 1 entrada e 2 saídas digitais, memória de 1 GB para coleta de dados, display IHM 320X240 TFT colour touch, grau de proteção IP 40, conexões 3/8" BSPP, dimensões de 104 x 126 x 34 mm e peso de 250 g.
Dispositivos Legais: RGI 1 e RGI 6 da NCM constante da TEC, aprovada pela Res. Gecex nº 272, de 2021, e da Tipi, aprovada pelo Dec. nº 11.158, de 2022; e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e pelas IN RFB nº 1.788, de 2018, nº 2.052, de 2021, e alterações posteriores.

MARCO ANTÔNIO RODRIGUES CASADO
Presidente da 5ª Turma

SOLUÇÃO DE CONSULTA Nº 98.023, DE 31 DE JANEIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM: 8479.89.99
Mercadoria: Aparelho de uso laboratorial para deposição por fotoresina de filmes finos de diversos tipos de materiais (com espessuras que podem variar de dezenas de nanômetros até centenas de micrômetros), a partir de solução, sobre superfícies rígidas ou flexíveis, de maneira controlada e precisa, pelo processo denominado comercialmente "Blade Coating".
Dispositivos Legais: RGI 1, RGI 6 e RGC 1, da NCM constante da TEC, aprovada pela Resolução Gecex nº 272, de 2021, e da Tipi, aprovada pelo Decreto nº 11.158, de 2022; e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e atualizada pela IN RFB nº 1.788, de 2018, e alterações posteriores.

LUIZ HENRIQUE DOMINGUES
Presidente da 4ª Turma

SOLUÇÃO DE CONSULTA Nº 98.024, DE 31 DE JANEIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM: 8543.90.10
Mercadoria: Placa de circuito impresso com componentes eletrônicos montados, própria para montagem de dispositivo para captação de dados de rede CAN Bus 2.0 de forma não invasiva, utilizado em veículos para coleta de parâmetros de operação tais como velocidade, rotação (rpm), pressão e temperatura do óleo, volume e consumo de combustível, etc., que serão tratados em outro dispositivo que a ele esteja conectado através de seu cabo, como um computador de bordo.
Dispositivos Legais: RGI 1, RGI 6 e RGC 1, da NCM constante da TEC, aprovada pela Resolução Gecex nº 272, de 2021, e da Tipi, aprovada pelo Decreto nº 11.158, de 2022; e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e atualizada pela IN RFB nº 1.788, de 2018, e alterações posteriores.

LUIZ HENRIQUE DOMINGUES
Presidente da 4ª Turma

SOLUÇÃO DE CONSULTA Nº 98.025, DE 31 DE JANEIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM: 8806.92.00
Ex Tipi: 01
Mercadoria: Veículo aéreo não tripulado de quatro rotores verticais, comercialmente denominado "drone" , próprio para ser pilotado remotamente ou para realizar voos programados sem a intervenção de um operador (missões), com peso máximo de decolagem de 1.487 g, dimensões de 248 x 248 mm (diagonal de 350 mm), velocidade máxima de 58 km/h e autonomia de voo de 27 minutos; contendo seis câmeras com sensores CMOS de 1/2,9 polegadas, capazes de capturar imagens coloridas e de banda estreita (RGB, azul, verde, vermelho, borda vermelha e infravermelho próximo no espectro de 840 nm ± 26 nm); apresentado numa maleta de transporte que inclui uma bateria, um controle remoto, um carregador de bateria, um cabo de alimentação, um cabo de comunicação USB 3.0 tipo-C e um adaptador USB; destinado a monitoramento agrícola por meio de imagens multiespectrais, que fornecem informações diversas sobre a saúde das plantas, seu crescimento, condições do solo, entre outras.
Dispositivos Legais: RGI 1 (Nota 1 do Capítulo 88) e RGI 6 da NCM constante da TEC, aprovada pela Res. Gecex nº 272, de 2021, e da Tipi, aprovada pelo Dec. nº 11.158, de 2022; RGC/Tipi 1; e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e pelas IN RFB nº 1.788, de 2018, nº 2.052, de 2021, e alterações posteriores.

MARCO ANTÔNIO RODRIGUES CASADO
Presidente da 5ª Turma

SOLUÇÃO DE CONSULTA Nº 98.026, DE 8 DE FEVEREIRO DE 2023

Assunto: Classificação de Mercadorias
Código NCM: 9609.10.00
Mercadoria: Sortido constituído de dois lápis grafite com borracha no topo, lapiseira automática, borracha de apagar em formato de coração e caneta marca-texto, acondicionados em uma única embalagem para venda a retalho.
Dispositivos legais: RGI 1 c/c RGI 3 c) e RGI 6 da NCM constante na TEC, aprovada pela Resolução Gecex nº 272, de 2021, e na Tipi, aprovada pelo Decreto nº 11.158, de 2022, e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e atualizadas pelas IN RFB nº 1.788, de 2018, e nº 2.052, de 2021, e alterações posteriores.

CLAUDIA ELENA FIGUEIRA CARDOSO NAVARRO
Presidente do Comitê

SOLUÇÃO DE CONSULTA Nº 98.027, DE 8 DE FEVEREIRO DE 2023

Assunto: Classificação de Mercadorias
Reforma de ofício a Solução de Consulta nº 14 - SRRF09/Diana, de 15 de fevereiro de 2012.
Código NCM: 8537.62.59
Mercadoria: Gabinete, denominado Case de Comando, que se destina a alojar uma câmera de televisão para circuito interno cuja função é executar os movimentos horizontais da câmera, que são denominados PAN, os verticais, TILT, e o controle para aproximação ou afastamento, ZOOM, daí a origem abreviada do seu nome comumente conhecido como "controlador PTZ". Visa, ainda, transmitir as imagens captadas pela câmera através de rede Ethernet com compactação de vídeo H.264 ou vídeo composto, além de permitir a entrada e saída de áudio. Proporciona, também, o controle remoto da câmera via RS-485 ou Ethernet. O produto é apresentado sem a câmera de TV. Em sua embalagem acompanham um par de luvas, um CD contendo manual, uma pequena lixa com óleo lubrificante utilizado para proteção dos parafusos, um cabo com conectores de 1.500 mm e um suporte de fixação feito em alumínio.
Dispositivos legais: RGI 1 c/c 3 c), RGI 6 e RGC 1 da NCM constante da TEC, aprovada pela Res. Gecex nº 272, de 2021, e da Tipi, aprovada pelo Dec. nº 11.158, de 2022, e subsídios extraídos das Nesh, aprovadas pelo Dec. nº 435, de 1992, e atualizadas pelas IN RFB nº 1.788, de 2018, e nº 2.052, de 2021, e alterações posteriores.

CLAUDIA ELENA FIGUEIRA CARDOSO NAVARRO
Presidente do Comitê

SUBSECRETARIA-GERAL DA RECEITA FEDERAL DO BRASIL
SUPERINTENDÊNCIA REGIONAL DA RECEITA FEDERAL DO BRASIL 1ª REGIÃO FISCAL
DELEGACIA DA RECEITA FEDERAL DO BRASIL EM BRASÍLIA
EQUIPE DE FISCALIZAÇÃO DE IPI, PIS/COFINS E IOF (EFI 1)

ATO DECLARATÓRIO EXECUTIVO DRF/BSA Nº 2, DE 15 DE FEVEREIRO DE 2023

Concede Registro Especial - Papel Imune

O AUDITOR FISCAL DA RECEITA FEDERAL DO BRASIL, integrante da Equipe de Fiscalização de IPI, PIS/COFINS e IOF (EFI 1), DRF BSA/DF, em face ao disposto nos arts. 1º e 2º da Lei nº 11.945, de 04 de junho de 2009, bem como ao estabelecido na Instrução Normativa RFB nº 1.817, de 24 de julho de 2018, e o que consta do processo nº 10265.023483/2023-68, declara:
Art. 1º Fica concedido o seguinte Registro Especial de Papel Imune para atividade de Gráfica (GF):
I - Registro Especial nº GP-01101/00265
II - Beneficiário: GRÁFICA E EDITORA CM LTDA
III - CNPJ: 44.867.868/0002-07
IV - Domicílio fiscal: Setor SIBS Quadra 02 Conjunto B Lote 10, S/N, Núcleo Bandeirante, Brasília - DF, CEP 71730-202
Art 2º O Registro Especial é válido pelo prazo de 3 (três) anos, a partir da data de publicação do presente Ato Declaratório Executivo, renovável pelo mesmo período, conforme art. 30 da Instrução Normativa RFB nº 1.817, de 24 de julho de 2018.
Art. 3º O contribuinte está obrigado ao cumprimento da legislação tributária em vigor e alterações posteriores, envolvendo operações com o papel destinado à impressão de livros, jornais e periódicos, em especial dos requisitos e exigências da Lei nº 11.945, de 04 de junho de 2009 e da Instrução Normativa RFB nº 1.817, de 24 de julho de 2018.
Art. 4º O não cumprimento das obrigações tributárias de que trata a IN RFB nº 1.817/2018, estabelecidos para a concessão do presente registro poderá, sem prejuízo das demais sanções cabíveis, ocasionar: a) o cancelamento do registro; b) a aplicação das penalidades previstas nos incisos I, II e § único, art. 17 da supracitada IN; c) poderá ser aplicado o regime especial de fiscalização previsto no art. 33 da Lei nº 9.430, de 27 de dezembro de 1996, uma vez configurada hipótese de crime contra a ordem tributária prevista no art. 2º da Lei nº 8.137, de 1990.
Art. 5º Este Ato Declaratório entra em vigor na data de sua publicação.

LUÍS CARLOS COIORIAN

ATO DECLARATÓRIO EXECUTIVO DRF/BSA Nº 4, DE 16 DE FEVEREIRO DE 2023

Concede Registro Especial - Papel Imune

O Auditor Fiscal da Receita Federal do Brasil, integrante da Equipe de Fiscalização de IPI, PIS/COFINS e IOF (EFI 1), DRF BSA/DF, em face ao disposto nos arts. 1º e 2º da Lei nº 11.945, de 04 de junho de 2009, bem como ao estabelecido na Instrução Normativa RFB nº 1.817, de 24 de julho de 2018, e o que consta do processo nº 10265.050225/2023-54, declara:
Art. 1º Fica concedido o seguinte Registro Especial de Papel Imune para atividade de Importador (IP):

Figura B.7: Exemplo de Diário Oficial contendo formatação de coluna dupla, porém distinta do observado nas Figuras B.1, B.2 e B.6 (Diário Oficial da União).

Protocolo: 2023000820577

Assunto: Portaria
Expediente: 000006-1285/23-0

Portarias - Portaria nº 14/2023-DG/IGP/SSP

O Diretor-Geral Adjunto do Instituto-Geral de Perícias, no uso de suas atribuições legais, tendo em vista as informações contidas na Demanda - Carteira de Identidade com erro de CPF * Protocolo nº 243814/0168, oriunda da Ouvidoria Setorial da Secretaria da Segurança Pública, noticiando possível irregularidade ou descumprimento de deveres funcionais por servidor lotado no Departamento de Identificação, Resolve: Instaurar sindicância administrativa nos termos do artigo 201 da Lei Complementar nº 10.898/94, designando os corregedores Fernando Ferreira Ipar, Perito Criminal, ID nº 2935578/3, Ilton Alves Baltazar, Papiloscopista, ID nº 1803999/1 e Rodrigo Cesar da Silva, Perito Criminal, ID nº 3094880/2 para, sob a Presidência do primeiro, compor Comissão para apurar o relatado no documento acima mencionado. Cumpra-se e publique-se. Porto Alegre, 16 de fevereiro de 2023.

Maiquel Luis Santos,
Diretor-Geral Adjunto do IGP/SSP

Protocolo: 2023000820578

Assunto: Portaria
Expediente: 000007-1285/23-3

Portarias - Portaria nº 15/2023- DG/IGP/SSP

O Diretor-Geral Adjunto do Instituto-Geral de Perícias, no uso de suas atribuições legais, tendo em vista as informações encaminhadas pelo Assistente Social da Penitenciária Estadual de Arroio dos Ratos, noticiando possível descumprimento de deveres funcionais por servidores lotados no Posto de Identificação de São Jerônimo, Resolve: Instaurar sindicância administrativa nos termos do artigo 201 da Lei Complementar nº 10.898/94, designando os corregedores Rodrigo Cesar da Silva, Perito Criminal, ID nº 3094880/2, Fernando Ferreira Ipar, Perito Criminal, ID nº 2935570/3 e Ilton Alves Baltazar, Papiloscopista, ID nº 1803999/1 para, sob a Presidência do primeiro, compor Comissão para apurar o relatado no documento acima mencionado. Cumpra-se e publique-se. Porto Alegre, 16 de fevereiro de 2023.

Maiquel Luis Santos,
Diretor-Geral Adjunto do IGP/SSP

Protocolo: 2023000820579

Assunto: Portaria
Expediente: 000008-1285/23-6

Portarias - Portaria nº 16/2023-DG/IGPSSP

O Diretor-Geral Adjunto do Instituto-Geral de Perícias, no uso de suas atribuições legais, tendo em vista a notícia de uso indevido do Sistema Consultas Integradas por servidor lotado no Posto Médico-Legal de Lajeado, Resolve: Instaurar sindicância administrativa nos termos do artigo 201 da Lei Complementar nº 10.898/94, designando os corregedores Ilton Alves Baltazar, Papiloscopista, ID nº 1803999/1, Fernando Ferreira Ipar, Perito Criminal, ID nº 2935570/3 e Rodrigo Cesar da Silva, Perito Criminal, ID nº 3094880/2 para, sob a Presidência do primeiro, compor Comissão para apurar o fato acima mencionado. Cumpra-se e publique-se. Porto Alegre, 16 de fevereiro de 2023.

Maiquel Luis Santos,
Diretor-Geral Adjunto do IGP/SSP

CORPO DE BOMBEIROS MILITAR DO RS

EDUARDO ESTEVAM CAMARGO RODRIGUES - CORONEL QOEM
Rua Silva Só, 300
Porto Alegre / RS / 90610-270

Gabinete do Comando Geral

EDUARDO ESTEVAM CAMARGO RODRIGUES - CORONEL QOEM

Convênios

Protocolo: 2023000820580

Assunto: Termo de Cooperação, Protocolo de Intenção
Expediente: 22/1200-0000826-6

Convênios - Termo de Cooperação, Protocolo de Intenção

Apostilamento ao Termo de Cooperação FPE nº 3292/2022

CONCEDENTE: Corpo de Bombeiros Militar do Estado do Rio Grande do Sul; CONVENENTE: Brigada Militar do Estado do Rio

Figura B.8: Exemplo de Diário Oficial contendo formatação de coluna simples, contendo diversos elementos de layout (Diário Oficial do Estado do Rio Grande do Sul).

22 ANO XLVIII - Nº 140 - PARTE I
SEGUNDA-FEIRA - 1 DE AGOSTO DE 2022
DIÁRIO OFICIAL
PODER EXECUTIVO

SECRETARIA DE ESTADO DE ADMINISTRAÇÃO PENITENCIÁRIA

EDITAL

DISPÕE SOBRE A CONVOCAÇÃO DE 01 (UM) CANDIDATO DO SEXO MASCULINO POR DECISÃO JUDICIAL APROVADO NA PROVA OBJETIVA CONFORME RELAÇÃO NOMINAL PUBLICADA NO D.O. DE 11 DE JUNHO DE 2012, E APROVADO NO EXAME DE APTIDÃO FÍSICA, PARA PROVIMENTO DE VAGAS NO CARGO DE INSPETOR DE SEGURANÇA E ADMINISTRAÇÃO PENITENCIÁRIA NA CLASSE INICIAL, EM CONCURSO PÚBLICO REALIZADO NO ANO DE 2012, OBSERVADO A ORDEM DE CLASSIFICAÇÃO PARA A REALIZAÇÃO DAS DEMAIS ETAPAS DO CERTAME, ESTABELECIDAS NO EDITAL PUBLICADO NO D.O. DE 01 DE MARÇO DE 2012.

A SECRETÁRIA DE ESTADO DE ADMINISTRAÇÃO PENITENCIÁRIA, em cumprimento à decisão judicial abaixo referenciada em trâmite perante o Tribunal de Justiça do Estado do Rio de Janeiro, devidamente orientada pela Procuradoria Geral do Estado, CONVOCA o candidato para o EXAME PSICOTÉCNICO do concurso em epígrafe, que será realizado no dia 11 de agosto de 2022, às 14h00min (devendo chegar com antecedência mínima de 30min), na Superintendência Central de Perícias Médicas e Saúde Ocupacional - SPMSO/SES, localizada na Rua Silva Jardim, nº 31, 3º andar, Centro (Praça Tiradentes), Rio de Janeiro - RJ, conforme relação abaixo:

MASCULINO

Nº	INSC	NOME	NOTA	PROCESSO
01	7139994	JOSÉ CARLOS SOUSA GUIMARÃES	86	Ação nº 0060163-06.2022.8.19.0001 SEI-140001/016631/2022

Secretaria de Estado de Defesa Civil

SECRETARIA DE ESTADO DE DEFESA CIVIL
CORPO DE BOMBEIROS MILITAR DO ESTADO DO RIO DE JANEIRO

EXTRATO DE TERMO

INSTRUMENTO: Termo de Ajuste de Contas nº 29/2020.
PARTES: O Estado do Rio de Janeiro, através do Corpo de Bombeiros Militar do Estado do Rio de Janeiro, e a empresa ZILLEO COPY COMERCIO E SERVIÇOS LTDA - EPP.
OBJETO: Reconhece que a empresa prestou serviços de impressão com fornecimento em regime de locação e instalação de equipamentos com tecnologia monocromática, juntamente com serviços de monitoramento, contabilização e manutenção corretiva "on-site", bem como o fornecimento de todos os suprimentos, incluindo papel, sem cobertura contratual válida.
VALOR: R$ 437.770,00 (quatrocentos e vinte e sete mil setecentos e setenta reais).
QUITAÇÃO: Ampla, geral, irrevogável e completa após o pagamento.
DATA DA ASSINATURA: 29/07/2022.
FUNDAMENTO: Processo nº SEI-270003/000059/2020.

SECRETARIA DE ESTADO DE DEFESA CIVIL
CORPO DE BOMBEIROS MILITAR DO ESTADO DO RIO DE JANEIRO

EXTRATO DE TERMO

INSTRUMENTO: 2º TERMO DE APOSTILAMENTO AO CONTRATO Nº 31/2019
PARTES: O Estado do Rio de Janeiro, através do Corpo de Bombeiros Militar do Estado do Rio de Janeiro, e a empresa KIARGOS SERVIÇOS E FACILITY LTDA.
OBJETO: Apostilar o Contrato nº 31/2019 no valor de R$ 197.146,83 (cento e oitenta e sete mil, cento e quarenta e seis reais e oitenta e cinco centavos), para fins de repactuação de preços, com base na CONVENÇÃO COLETIVA DE TRABALHO 2020/2021, do SINDICATO DOS TRABALHADORES NAS EMPRESAS DE REFEIÇÕES COLETIVAS REFEIÇÕES RAPIDAS (FASTFOOD) E AFINS DO ESTADO DO RIO DE JANEIRO - SINDIRREFEICOES-RJ e SINDICATO DAS EMPR. DE REF. COLET. DEST DO RIO DE JANEIRO, REGISTRADA NO MTE RJ030701/2020, com vigência de 01º de janeiro de 2020 a 31 de outubro de 2021 e a data-base da categoria em 01º de janeiro e na CONVENÇÃO COLETIVA DE TRABALHO 2021/2022, DO SINDICATO DAS EMPRESAS DE ASSEIO E CONSERVAÇÃO ESTADO DO RIO DE JANEIRO, REGISTRADA NO MTESGDR D Nº RJ000713/2021, MR013859/2021, com vigência de 1º/03/2021 a 28/02/2023, data-base da categoria em 01º de março, em razão do reajuste concedido à categoria profissional e do reajuste do valor do auxílio alimentação ou refeição.
VIGÊNCIA: A contar da data de publicação deste instrumento no DOERJ.
VALOR TOTAL: O valor global do contrato passa de R$ 2.822.736,12 (dois milhões, oitocentos e vinte e dois mil, setecentos e trinta e seis reais e doze centavos) para R$ 3.009.885,97 (três milhões, nove mil, oitocentos e oitenta e cinco reais e noventa e sete centavos).
DATA DA ASSINATURA: 01/04/2022.
FUNDAMENTO: Processo nº SEI-270003/001884/2021.
*Republicado por ter saído com incorreção no original publicado no D.O. de 03.06.2022.

SECRETARIA DE ESTADO DE DEFESA CIVIL

AVISOS

PREGÃO ELETRÔNICO Nº 31/22-RJ.
OBJETO: Registro de preços para eventual aquisição de Caixa Térmica Tipo Hotbox.
DATA DE ABERTURA: 12/08/2022, às 09h30min.
DATA ETAPA DE LANCES: 12/08/2022, às 09h
PROCESSO Nº SEI-270003/000161/2022.

PREGÃO ELETRÔNICO Nº 46/22.
OBJETO: Registro de preços para a aquisição de medicamentos.
DATA DE ABERTURA: 15/08/2022, às 09h
DATA ETAPA DE LANCES: 15/08/2022, às 09h30min
PROCESSO Nº SEI-270003/000852/2021.

Os Editais encontram-se à disposição dos interessados nos sites: www.comprasnet.rj.gov.br ou www.cbmerj.rj.gov.br/licitacoes, podendo ser retirados, de forma impressa, na Coordenação de Licitações e Contratos/DGAF/SEDEC, sito à Praça da República, nº 45 - Centro - RJ, de 2ª a 5ª feira, das 08:00 às 17:00 horas, e 6ª feira, das 08:00 às 12:00 horas. Informações pelos Tels. (21) 3333-3084/3333-3085 ou pelo e-mail: pregaoeletronico@cbmerj.rj.gov.br

CORPO DE BOMBEIROS MILITAR DO ESTADO DO RIO DE JANEIRO

EDITAL

O **COMANDANTE-GERAL DO CORPO DE BOMBEIROS MILITAR DO ESTADO DO RIO DE JANEIRO**, no uso de suas atribuições, **CONVOCA**: o ex-militar **ROBSON PIRES DA SILVA**, CPF 098.565.177-61, para comparecer à Diretoria-Geral de Finanças (DGF), sito à Praça da República, nº 45, Centro, Rio de Janeiro, para tratar de assunto de seu interesse, de segunda-feira a quinta-feira, das 09:00h às 15:00h, impreterivelmente, até o dia 26 (vinte e seis) de agosto de 2022. Processo nº SEI-270115/000491/2020.

Secretaria de Estado de Saúde

SECRETARIA DE ESTADO DE SAÚDE

EXTRATO DE TERMO

INSTRUMENTO: Termo de Ajuste de Contas nº 059/2022.
PARTES: Secretaria de Estado de Saúde e BRISTOL MYERS SQUIBB FARMACÊUTICA LTDA.
OBJETO: Prestação do Serviço de Fornecimento de Medicamento para atender a Superintendência de Assistência Farmacêutica e Insumos Estratégicos, no período de junho de 2018.
VALOR: R$ 79.789,00 (setenta e nove mil setecentos e oitenta e oito reais).
DATA DA ASSINATURA: 28/07/2022.
FUNDAMENTO LEGAL: Lei Federal nº 8.666 de 21/06/1993, pela Lei Estadual nº 287, de 04/12/1979, e seu Regulamento, aprovado pelo Decreto Estadual nº 3.149, de 28/04/1980.
PROCESSO Nº SEI-060001/011602/2018.

SECRETARIA DE ESTADO DE SAÚDE

EXTRATO DE TERMO

INSTRUMENTO: TERMO DE CREDENCIAMENTO Nº 012/2022.
PARTES: ESTADO DO RIO DE JANEIRO pela SECRETARIA DE ESTADO DE SAÚDE e HOSPITAL VIVER MAIS LTDA.
OBJETO: Credenciamento pela Secretaria de Estado de Saúde da unidades da saúde do setor privado, destinadas ao atendimento a pacientes que necessitam dos serviços de cirurgia bariátrica por videolaparoscopia e acompanhamento pré e pós operatório multidisciplinar referenciados pela Secretaria de Estado de Saúde, através da Superintendência de Regulação.
VIGÊNCIA: 12 (doze) meses, contados a partir da assinatura do Termo de Credenciamento.
VALOR TOTAL: R$ 11.306.922,92 (onze milhões, trezentos e seis mil novecentos e vinte e dois reais e noventa e dois centavos).
FUNDAMENTO DO ATO: Lei nº 8.666 de 21 de junho de 1.993 e alterações, Lei Estadual nº 287 de 04 de dezembro de 1979, Decretos nº 3.149/80 e 21.081/94 , com as respectivas alterações, além das demais disposições legais aplicáveis, cumprindo definido nas Portarias MS/GM nº 424 e 425/13, de 19 de março de 2013, que estabelecem regulamentos técnicos, normas e critério para o serviço da Assistência de Alta Complexidade ao indivíduo com obesidade, Resolução do Conselho Federal de Medicina - Resolução CFM nº 1.942/2010, publicada no D.O.U. de 12 de fevereiro de 2010, seção I, p. 72, além da Portaria MS nº 492, de 31 de Agosto de 2007.
DATA DA ASSINATURA: 27/07/2022.
PROCESSO Nº SEI-060001/020896/2021.

SECRETARIA DE ESTADO DE SAÚDE
SUBSECRETARIA EXECUTIVA
SUPERINTENDÊNCIA DE COMPRAS E LICITAÇÕES
COORDENAÇÃO DE LICITAÇÃO

AVISO

A COORDENAÇÃO DE LICITAÇÃO/SES torna pública a seguinte licitação:

PREGÃO ELETRÔNICO Nº 353/22.
OBJETO: Aquisição de resmas de 500 folhas de papel sulfite para impressão nos formatos (A3 e A4) e bobina de papel para plotter, para atender as necessidades do nível central da Secretaria de Estado de Saúde do Rio de Janeiro (SES-RJ) e unidades externas sob sua gestão, na forma do Termo de Referência (ANEXO 01).
ABERTURA DAS PROPOSTAS: 12/08/2022, às 10:00 horas.
ETAPA DE LANCES: 12/08/2022, às 10:00 horas.
PROCESSO Nº SEI-060001/000629/2022.

O edital encontra-se à disposição dos interessados nos sites: www.comprasnet.rj.gov.br , https://sei.fazenda.rj.gov.br e www.saude.rj.gov.br/licitacoes. Podendo também ser retirado de forma impressa, na Coordenação de Licitação, mediante a entrega de 01 (uma) resma de papel tamanho A4, sito à Rua México, Nº 128 - 9º andar, sala 905 - Centro - Rio de Janeiro - RJ, de 2ª a 6ª feira, das 10:00 às 16:00 hs. Informações pelo e-mail: licitacao@saude.rj.gov.br.

ADMINISTRAÇÃO VINCULADA

INSTITUTO VITAL BRAZIL S.A

EXTRATO DE INSTRUMENTO CONTRATUAL

INSTRUMENTO: Contrato nº 014/2022. **PARTES**: INSTITUTO VITAL BRAZIL S.A (IVB) e SODEXO PASS DO BRASIL SERVICOS E COMERCIO S.A. **OBJETO**: prestação de serviços de empresa especializada para fornecimento de auxílio alimentação/refeição, na forma de cartão magnético com chip, para atender as necessidades dos empregados públicos do Instituto Vital Brazil. **PRAZO**: 12 (doze) meses. **VALOR TOTAL**: R$ 1.634.273,40 (um milhão, seiscentos e cinquenta e quatro mil duzentos e setenta e três reais e quarenta centavos). **DATA DA ASSINATURA**: 29/07/2022. **FISCAIS**: Paulo Cesar da Silveira Sodré ID: 2687886-9 e Claudia Gouveia Mendes. ID: 4142746-7. **PROC. IVB Nº SEI-080005/000782/2022.**

SECRETARIA DE ESTADO DE SAÚDE
FUNDAÇÃO SAÚDE DO ESTADO DO RIO DE JANEIRO

EXTRATO DE INSTRUMENTO CONTRATUAL

INSTRUMENTO: Contrato nº 266/2022. PREGÃO ELETRÔNICO nº 22/2022. **PARTES**: Fundação Saúde do Estado do Rio de Janeiro e a empresa SISTEMA DE EMERGENCIA MOVEL DE BRASILIA LTDA. **OBJETO**: contratação de empresa para locação de veículos, tipo ambulâncias, para operacionalização do SAMU-192, para atender as demandas desta unidade em um período de 12 (doze) meses, conforme descrito no item 7 do Anexo 01- Termo de Referência doc. SEI 26416374. **VIGÊNCIA**: 12 (doze) meses, a contar de 01/08/2022. **VALOR TOTAL**: R$ 6.577.348,80 (seis milhões e quinhentos e setenta e sete mil e trezentos e quarenta e oito reais e oitenta centavos). **NOTA DE EMPENHO**: 2022NE04184 / 2022NE04185.**FUNDAMENTAÇÃO:** Lei nº 8.666, de 21 de junho de 1.993 e alterações, pela Lei Estadual nº 287, de 04 de dezembro de 1.979 e Decreto nº's 3.149, de 28 de abril de 1980, e 42.301, de 12 de fevereiro de 2010. DATA DA ASSINATURA: 29/07/2022. Parecer 203/2022 (doc. SEI 26892746) e Autorização do Ordenador de Despesa (doc. SEI 32578437). **PROCESSO Nº SEI-080007/001457/2021.**

SECRETARIA DE ESTADO DE SAÚDE
FUNDAÇÃO SAÚDE DO ESTADO DO RIO DE JANEIRO

EXTRATO DE INSTRUMENTO CONTRATUAL

INSTRUMENTO: CONTRATO Nº 243/2022. DISPENSA DE LICITAÇÃO Nº 190/2022. **PARTES**: Fundação Saúde do Estado do Rio de Janeiro e a empresa LIFE FARMA PRODUTOS MEDICOS EIRELI. **OBJETO**: AQUISIÇÃO DE MEDICAMENTOS (FOSFATO DISSÓDICO DE DEXAMETASONA) - item: 2, o que constitui itens fundamentais para a manutenção das atividades desenvolvidas pelas UPAs 24h geridas pela FSERJ, SAMU e das unidades HEM, HMHS, CEDI, HEMORIO, IDCAC, IEDE, HESM, IETAP, HECC, CPRJ, HEAN, IEDS, HEER e Hospital Doutor Hamilton Agostinho Vieira Castro - SEAP. **VIGÊNCIA**: 180 (cento e oitenta) dias, contados da data de assinatura. **VALOR TOTAL**: R$ 1.116.826,84 (um milhão, cento e dezessete mil, oitocentos e vinte e seis reais e oitenta e quatro centavos). **NOTA DE EMPENHO**: 2022NE04084. **FUNDAMENTAÇÃO:** art. 24, inciso IV da Lei nº 8.666, de 21 de junho de 1.993 e alterações, pela Lei Estadual nº 287, de 04 de dezembro de 1.979 e Decretos nº's 3.149, de 28 de abril de 1980, e 42.301, de 12 de fevereiro de 2010. **DATA DA ASSINATURA:** 28/07/2022. Parecer 144/2022 (34557847) e Autorização do Ordenador de Despesas (doc. SEI 36112580. **PROCESSO Nº SEI-080007/002667/2022.**

SECRETARIA DE ESTADO DE SAÚDE
FUNDAÇÃO SAÚDE DO ESTADO DO RIO DE JANEIRO

EXTRATO DE INSTRUMENTO CONTRATUAL

INSTRUMENTO: Contrato de COMODATO nº 014/2022. **PARTES**: Fundação Saúde do Estado do Rio de Janeiro e a empresa SAMTRONIC INDUSTRIA E COMERCIO LTDA. **OBJETO**: uso e título gratuito e com exclusividade, do bem: 177 BOMBAS INFUSORAS em sistema de COMODATO, para atender as demandas das unidades sob gestão da FSERJ. **VIGÊNCIA:** O prazo de vigência do Contrato de Comodato terá início a partir da data de sua publicação, em forma de extrato, no Diário Oficial do Estado, e se estenderá 06 (seis) meses, após o decurso do prazo de vigência final da referida Ata (19/12/2023) ou até o término do estoque dos insumos, podendo ser prorrogado, assim convinhado ao interesse público e por iniciativa das partes, através de termo aditivo. **FUNDAMENTO:** Processo Administrativo nº SEI-080007/005432/2021 e na ARP nº 060/2022-A, que se regerá pela legislação aplicável, em especial os artigos 579 a 585 do Código Civil, Lei Federal nº 8.666/93 e Lei Estadual nº 287/79. **DATA DA ASSINATURA**: 27/07/2022.

SECRETARIA DE ESTADO DE SAÚDE
FUNDAÇÃO SAÚDE DO ESTADO DO RIO DE JANEIRO

EXTRATOS DE ATAS DE REGISTRO DE PREÇOS

INSTRUMENTO: Ata de Registro de Preços nº 086/2022-A. PREGÃO ELETRÔNICO nº 55/2022. **PARTES**: Fundação Saúde do Estado do Rio de Janeiro e a empresa RALIC COMERCIO E DISTRIBUIÇÃO LTDA. **OBJETO**: (SUCCINATO SODICO DE METILPREDNISOLONA) - item: 5, item vital para a manutenção das atividades terapêuticas desenvolvidas pelas unidades de saúde sob gestão da Fundação Saúde do Estado do Rio de Janeiro - FSERJ. **VIGÊNCIA**: 12 (doze) meses, a contar da data da publicação do instrumento no DOERJ. **VALOR TOTAL REGISTRADO:** R$ 97.053,20 (noventa e sete mil setecentos e cinquenta três reais e vinte centavos). **FUNDAMENTO**: Lei Federal nº 8.666, de 21 de junho de 1993, Lei nº 10.520, de 17 de julho de 2002, do Decreto Estadual nº 46.751, de 27 de agosto de 2019, da Lei Estadual nº 287, de 4 de dezembro de 1979, do Decreto Estadual nº 3.149, de 28 de abril de 1980, e respectivas alterações, do instrumento convocatório, Parecer 823/2022 (Doc. 31238328) e Autorização do Ordenador de Despesa (Doc. SEI 34630365). **DATA DA ASSINATURA**: 28/07/2022. **PROCESSO Nº SEI-080007/000806/2021.**

INSTRUMENTO: Ata de Registro de Preços nº 086/2022-B. PREGÃO ELETRÔNICO nº 55/2022. **PARTES**: Fundação Saúde do Estado do Rio de Janeiro e a empresa NOVA LINEA COMERCIO DE PRODUTOS FARMACÊUTICOS LTDA. **OBJETO**: (BROMETO DE IPATROPIO + METILPREDNISOLONA SUCCINATO) - itens: 3 e 6, item vital para a manutenção das atividades terapêuticas desenvolvidas pelas unidades de saúde sob gestão da Fundação Saúde do Estado do Rio de Janeiro - FSERJ. **VIGÊNCIA**: 12 (doze) meses, a contar da data de publicação do instrumento no DOERJ. **VALOR TOTAL REGISTRADO:** R$ 236.549,60 (duzentos e trinta e seis mil quinhentos e quarenta e nove reais). **FUNDAMENTO**: Lei Federal nº 8.666, de 21 de junho de 1993, Lei nº 10.520, de 17 de julho de 2002, do Decreto Estadual nº 46.751, de 27 de agosto de 2019, da Lei Estadual nº 287, de 4 de dezembro de 1979, do Decreto Estadual nº 3.149, de 28 de abril de 1980, e respectivas alterações, do instrumento convocatório, Parecer 823/2022 (Doc. 31238328) e Autorização do Ordenador de Despesa (Doc. SEI 34630365). **DATA DA ASSINATURA**: 28/07/2022. **PROCESSO Nº SEI-080007/008866/2021.**

INSTRUMENTO: Ata de Registro de Preços nº 086/2022-C. PREGÃO ELETRÔNICO nº 55/2022. **PARTES**: Fundação Saúde do Estado do Rio de Janeiro e a empresa DIMASTER COMERCIO DE PRODUTOS HOSPITALARES LTDA. **OBJETO**: AQUISIÇÃO DE MEDICAMENTOS BRONCODILATADORES (SULFATO SALBUTAMOL) - item: 9, item vital para a manutenção das atividades terapêuticas desenvolvidas pelas unidades de saúde sob gestão da Fundação Saúde do Estado do Rio de Janeiro - FSERJ. **VIGÊNCIA**: 12 (doze) meses, a contar da data da publicação do instrumento no DOERJ. **VALOR TOTAL REGISTRADO:** R$ 134.561,00 (cento e trinta e quatro mil quinhentos e sessenta e um reais). **FUNDAMENTO**: Lei Federal nº 8.666, de 21 de junho de 1993, Lei nº 10.520, de 17 de julho de 2002, do Decreto Estadual nº 46.751, de 27 de agosto de 2019, da Lei Estadual nº 287, de 4 de dezembro de 1979, do Decreto Estadual nº 3.149, de 28 de abril de 1980, e respectivas alterações, do instrumento convocatório, Parecer 823/2022 (Doc. 31238328) e Autorização do Ordenador de Despesa (Doc. SEI 34630365). **DATA DA ASSINATURA**: 28/07/2022. **PROCESSO Nº SEI-080007/008866/2021.**

SECRETARIA DE ESTADO DE SAÚDE
FUNDAÇÃO SAÚDE DO ESTADO DO RIO DE JANEIRO

EXTRATOS DE ATAS DE REGISTRO DE PREÇOS

INSTRUMENTO: Ata de Registro de Preços nº 087/2022-A. PREGÃO ELETRÔNICO nº 24/2022. **PARTES**: Fundação Saúde do Estado do Rio de Janeiro e a empresa LEMAN MEDICAMENTOS E CIA LTDA. **OBJETO**: AQUISIÇÃO DE MEDICAMENTOS CARDIOVASCULARES (NITRATO DE ISOSSORBIDA) - item: 6, 8 e 11, para atender as demandas das unidades sob gestão da FSERJ. **VIGÊNCIA**: 12 (doze) meses, a contar da data de publicação do instrumento no DOERJ.

Figura B.9: Exemplo de Diário Oficial contendo formatação de coluna tripla, divergente dos demais diários exemplificados, além de elementos de coluna simples ao topo da página (Diário Oficial do Estado do Rio de Janeiro).

Apêndice C

Distribuição de Probabilidade dos Registros Classificados pelos Modelos Obtidos

Com o objetivo de prover o detalhamento da distribuição das probabilidades atribuídas a cada registro do conjunto de testes, foram obtidos os gráficos da Estimativa de Densidade de Kernel (EDK)[1]. Assim, torna-se possível verificar a tendência à esquerda, centro ou direita das distribuições referentes às classificações negativas ou positivas. As seções C.1 e C.2 contém os gráficos obtidos a partir dos modelos cujas saídas são probabilidades, sendo referentes os modelos balanceados e desbalanceados, respectivamente.

Observou-se que os modelos com melhores resultados apresentaram uma maior densidade de registros localizados à direita para a classe positiva. Essa observação reforça a conclusão da alta confiança das classificações dos modelos obtidos, o que levou a altos limiares otimizados de classificação. Os gráficos obtidos também evidenciaram os resultados inferiores, a exemplo das Figuras C.1, C.3 e C.14, com distribuições tendendo ao centro ou à esquerda.

[1] https://seaborn.pydata.org/tutorial/distributions.html#tutorial-kde

C.1 Distribuição para modelos Balanceados

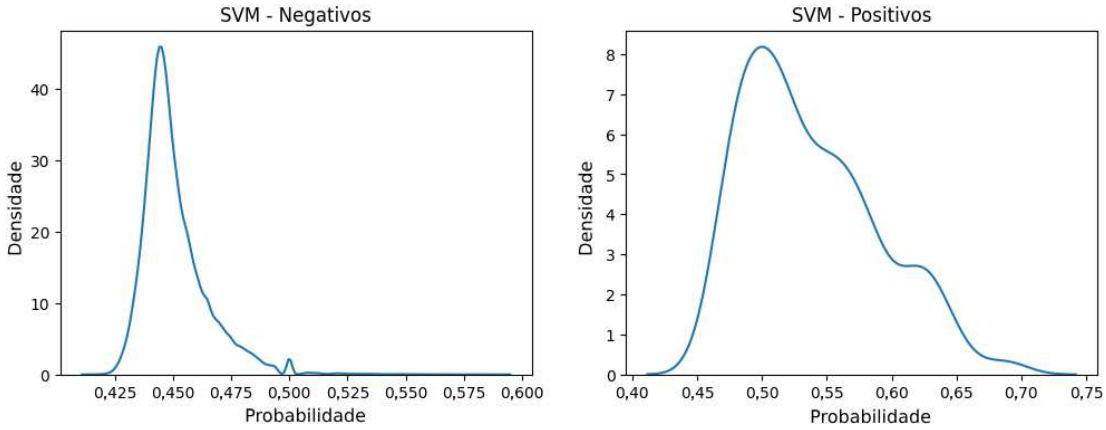


Figura C.1: Gráfico da EDK para os resultados do modelo SVM balanceado.

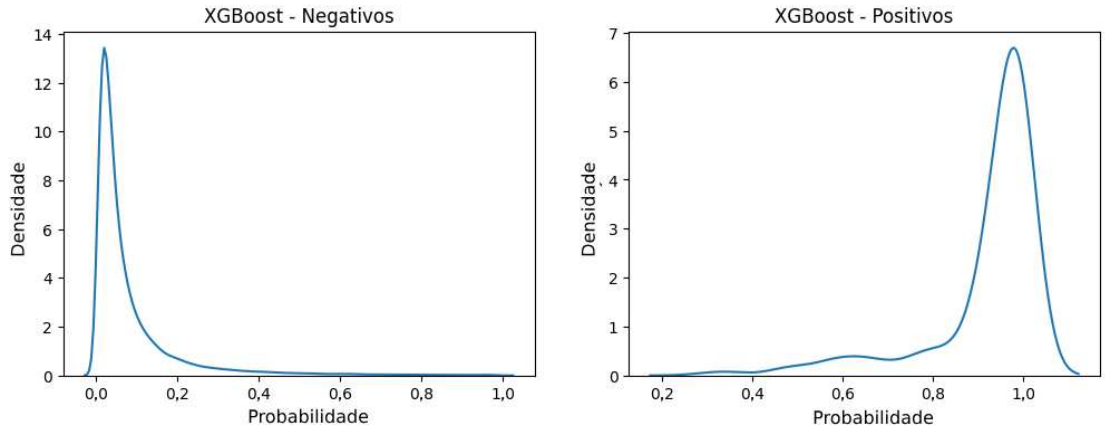


Figura C.2: Gráfico da EDK para os resultados do modelo XGBoost balanceado.

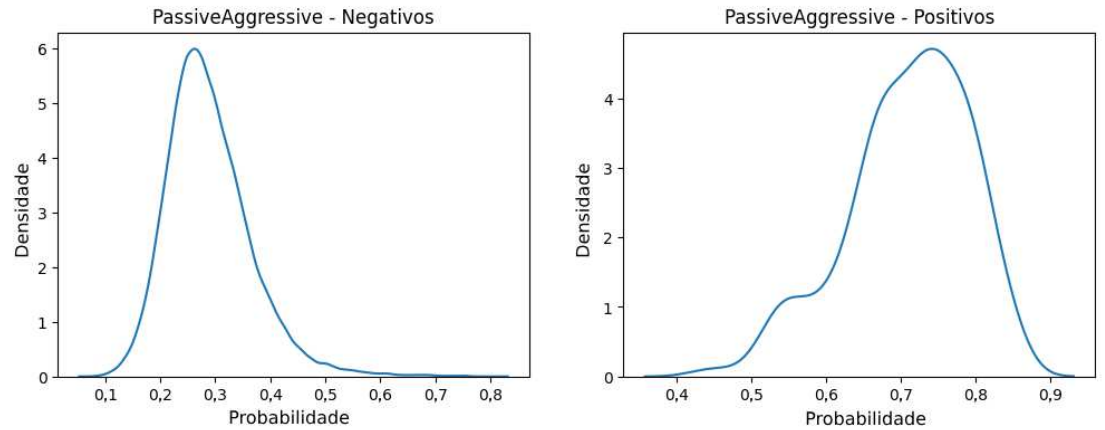


Figura C.3: Gráfico da EDK para os resultados do modelo *Passive Aggressive* balanceado.

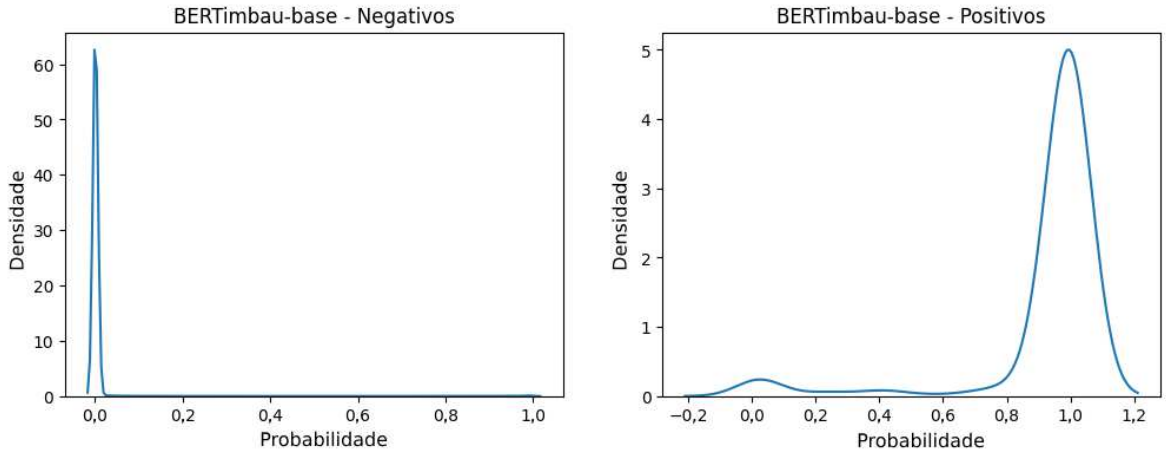


Figura C.4: Gráfico da EDK para os resultados do modelo BERTimbau-base balanceado.

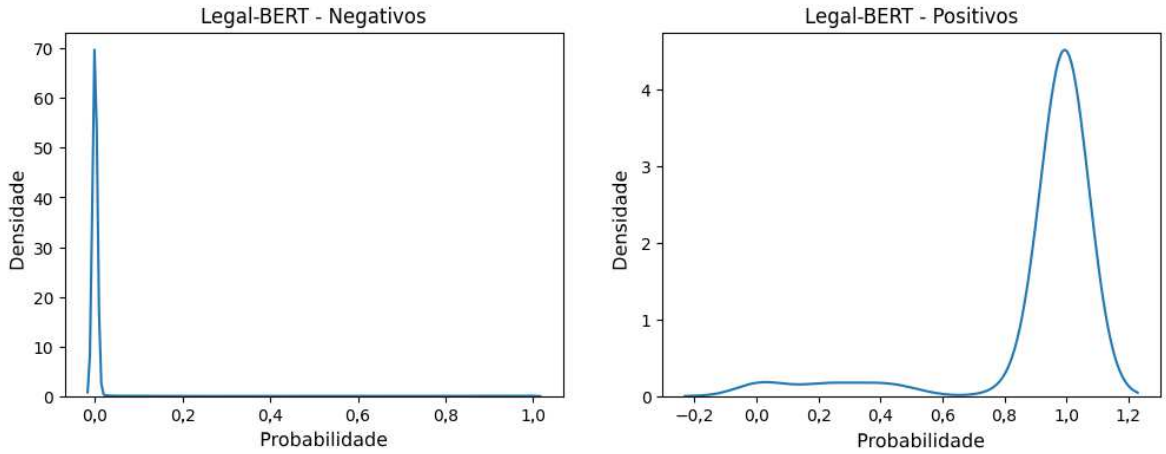


Figura C.5: Gráfico da EDK para os resultados do modelo Legal-BERT balanceado.

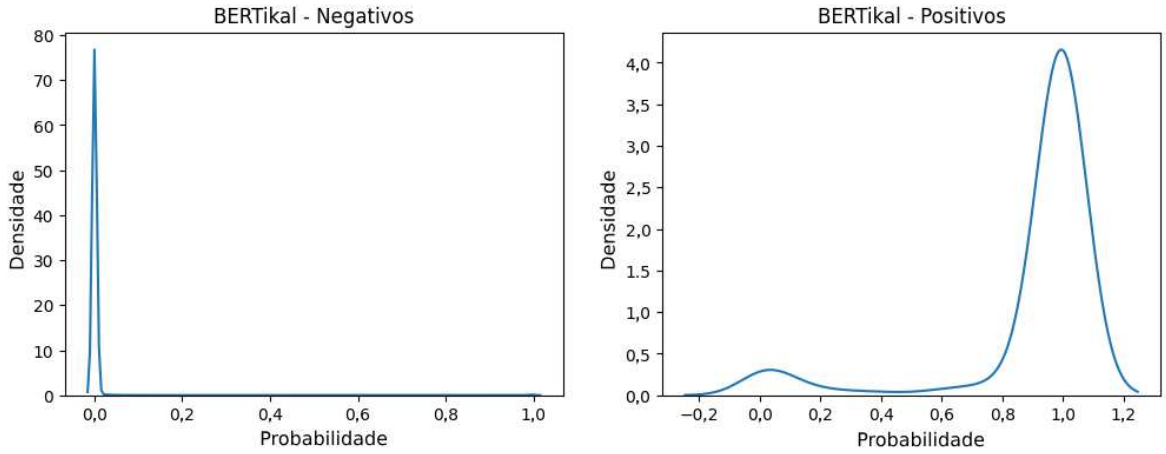


Figura C.6: Gráfico da EDK para os resultados do modelo BERTikal balanceado.

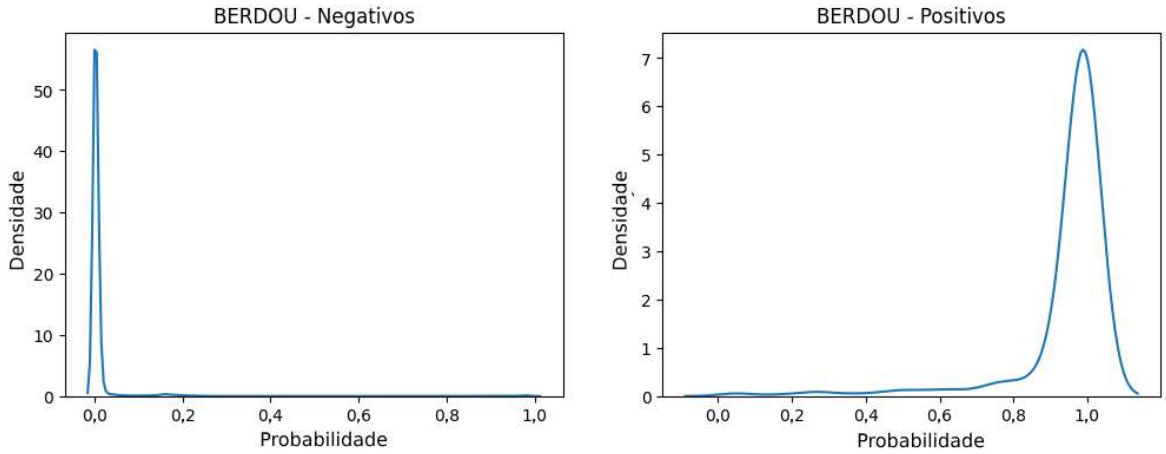


Figura C.7: Gráfico da EDK para os resultados do modelo BERDOU balanceado.

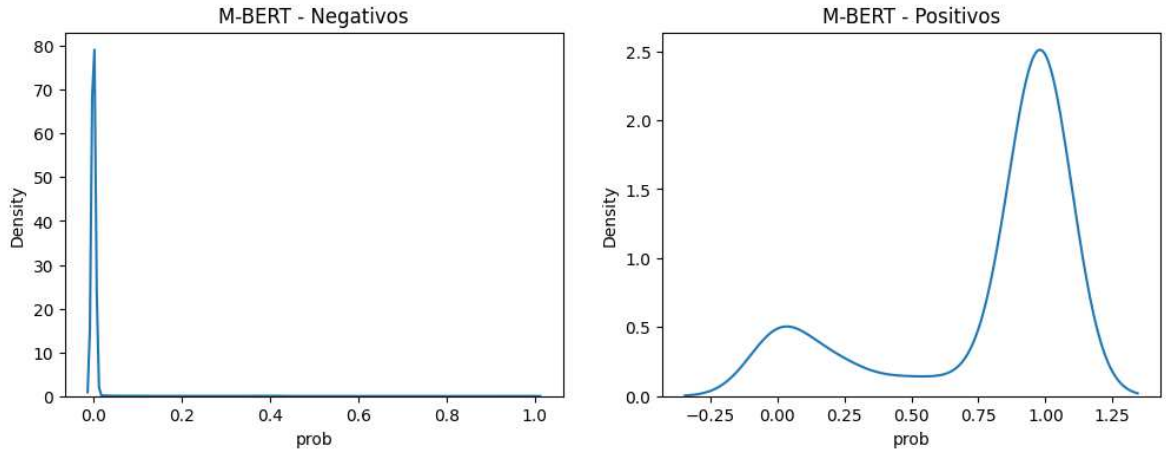


Figura C.8: Gráfico da EDK para os resultados do modelo M-BERT balanceado.

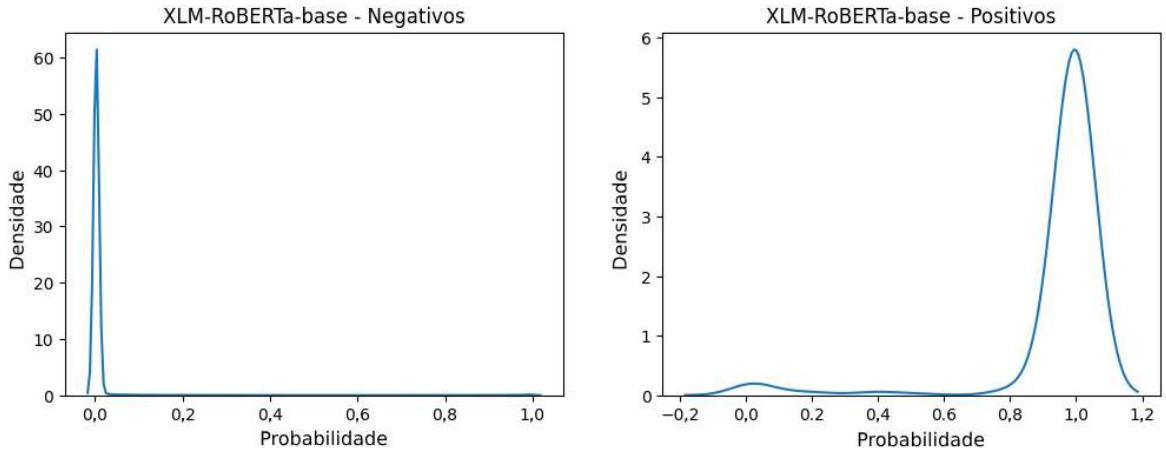


Figura C.9: Gráfico da EDK para os resultados do modelo XLM-RoBERTa-base balanceado.

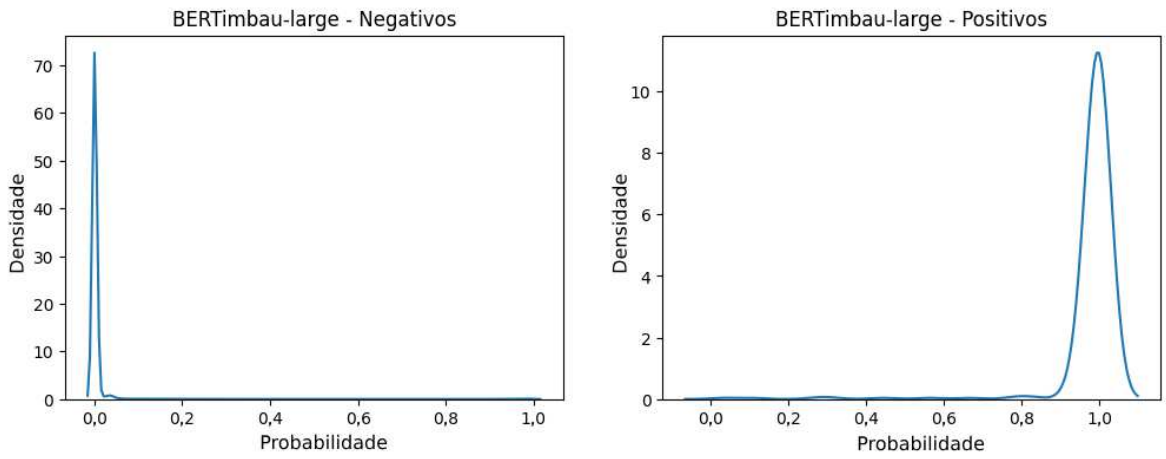


Figura C.10: Gráfico da EDK para os resultados do modelo BERTimbau-large balanceado.

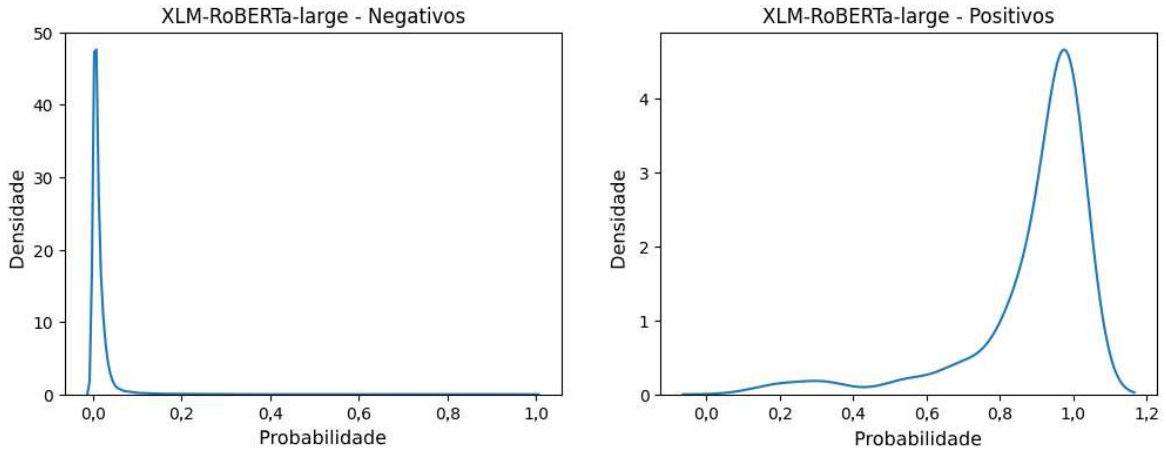


Figura C.11: Gráfico da EDK para os resultados do modelo XLM-RoBERTa-large balanceado.

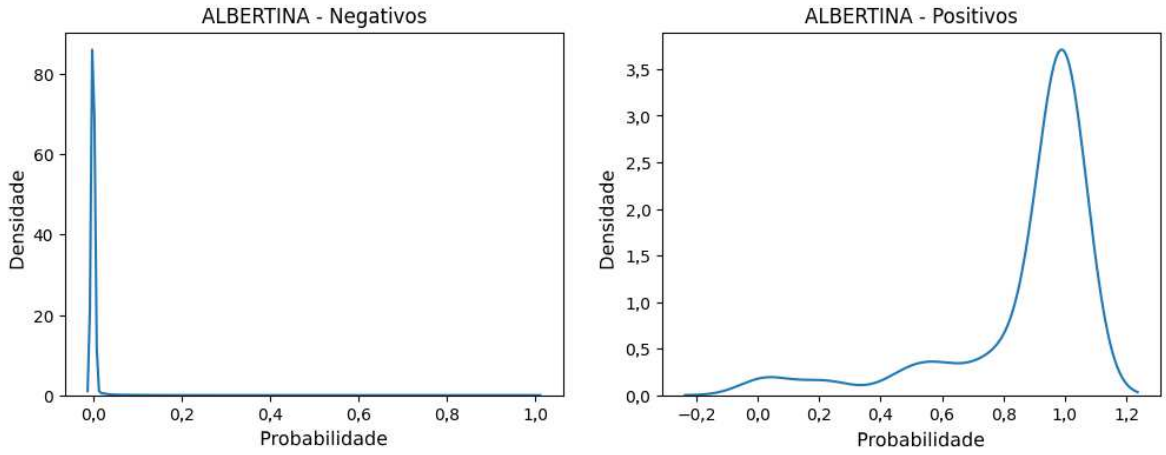


Figura C.12: Gráfico da EDK para os resultados do modelo ALBERTINA balanceado.

C.2 Distribuição para modelos Desbalanceados

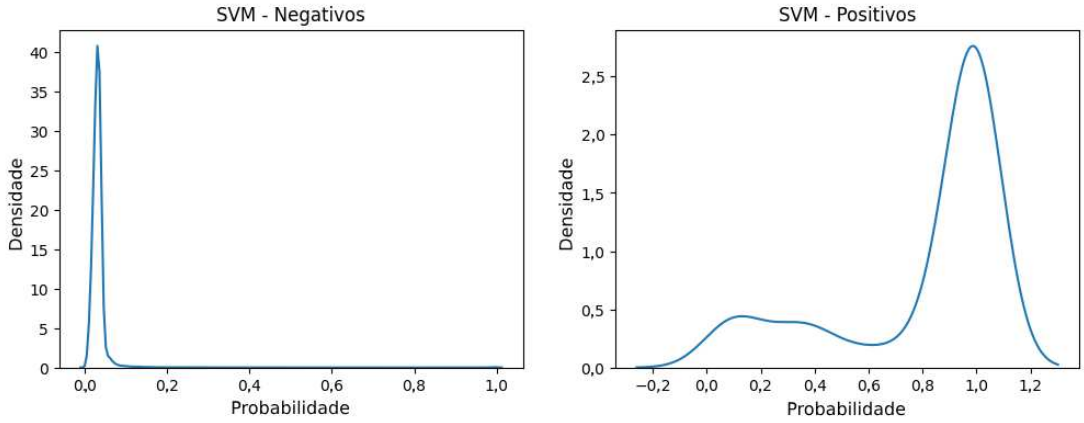


Figura C.13: Gráfico da EDK para os resultados do modelo SVM desbalanceado.

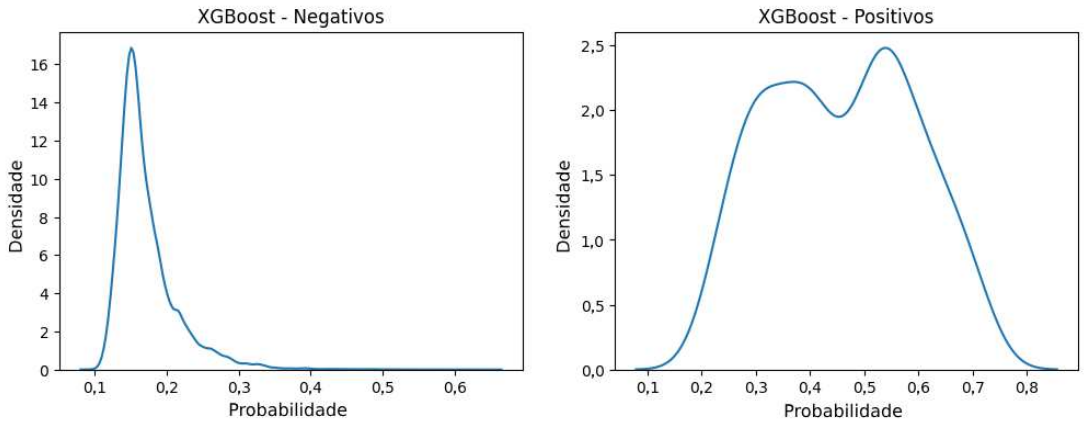


Figura C.14: Gráfico da EDK para os resultados do modelo XGBoost desbalanceado.

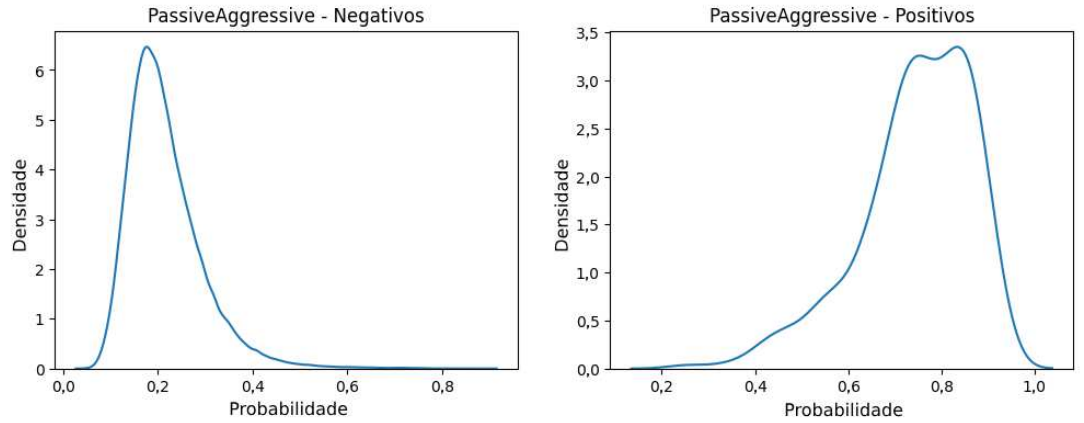


Figura C.15: Gráfico da EDK para os resultados do modelo *Passive Aggressive* desbalanceado.

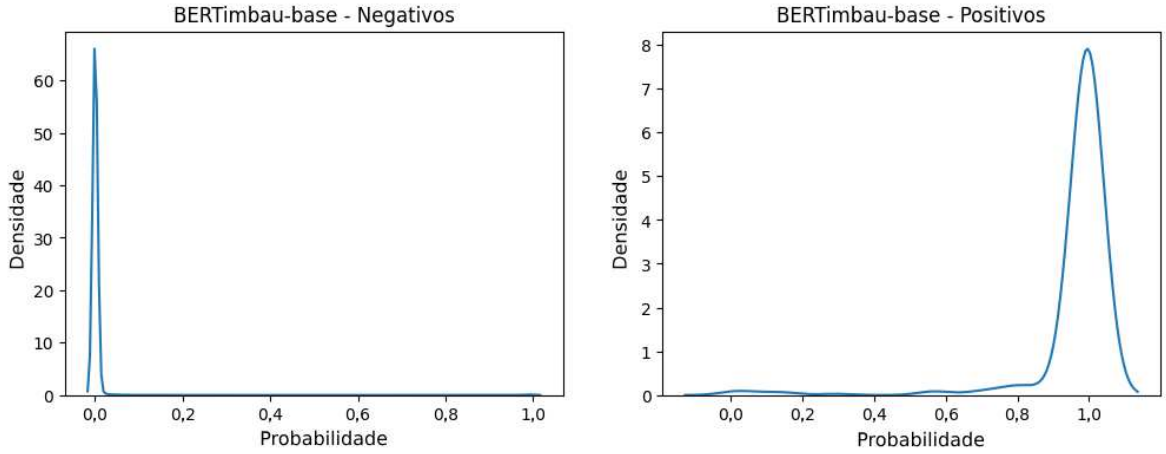


Figura C.16: Gráfico da EDK para os resultados do modelo BERTimbau-base desbalanceado.

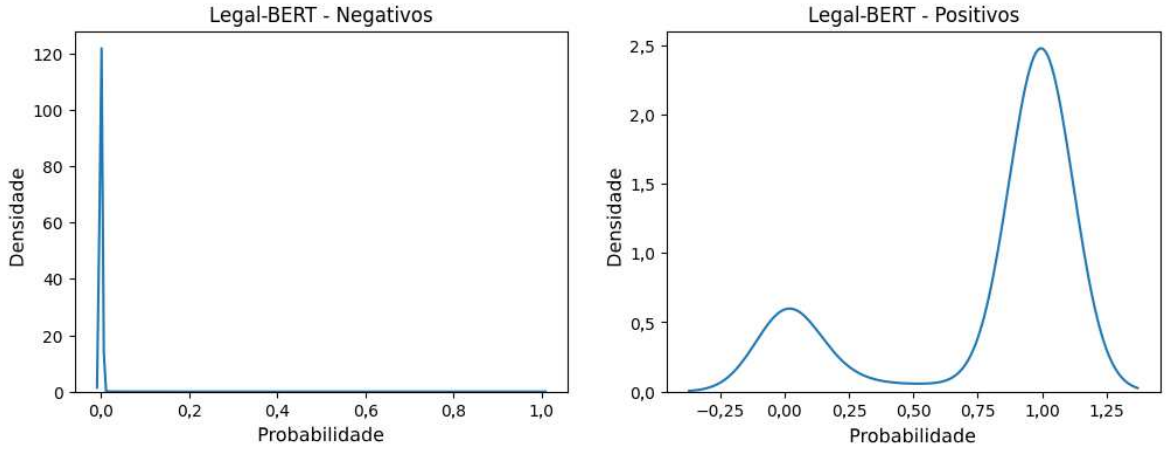


Figura C.17: Gráfico da EDK para os resultados do modelo Legal-BERT desbalanceado.

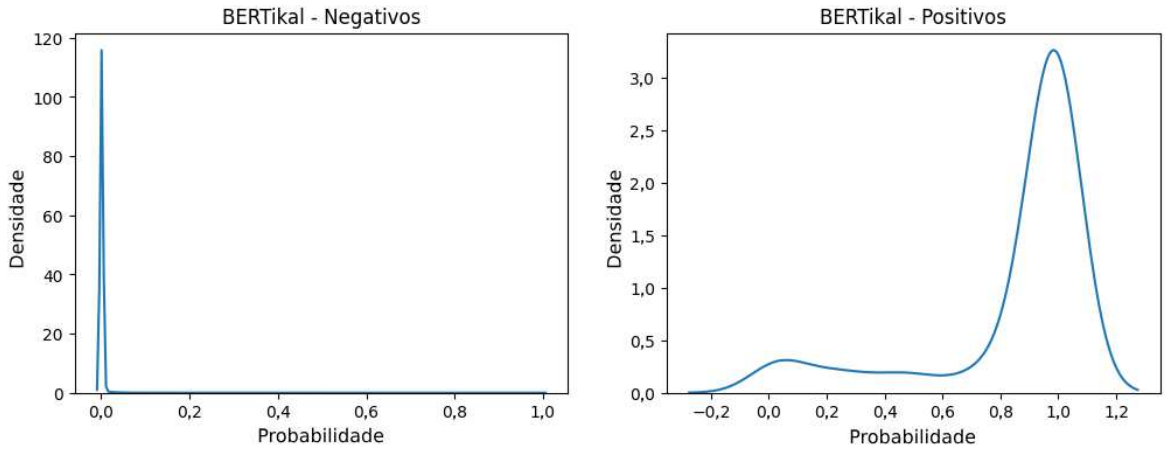


Figura C.18: Gráfico da EDK para os resultados do modelo BERTikal desbalanceado.

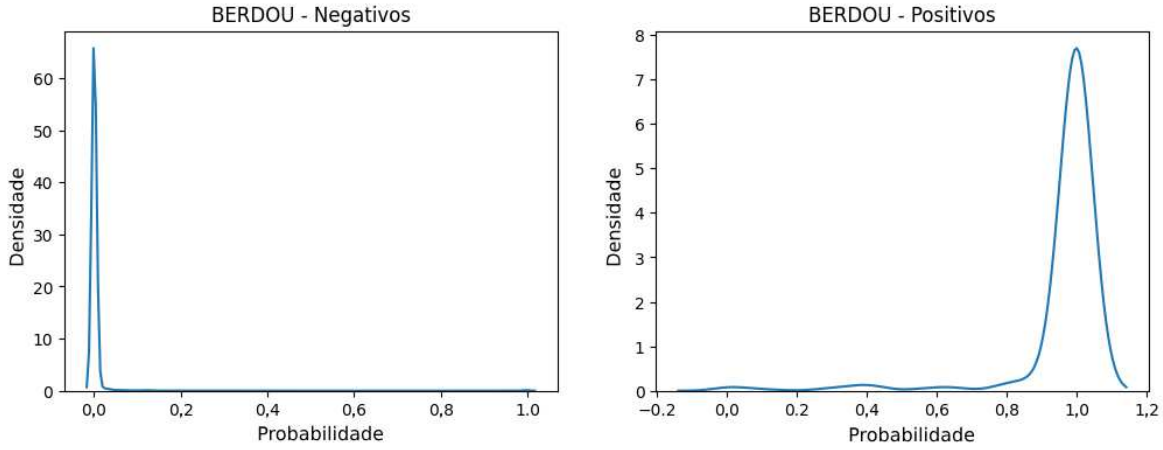


Figura C.19: Gráfico da EDK para os resultados do modelo BERDOU desbalanceado.

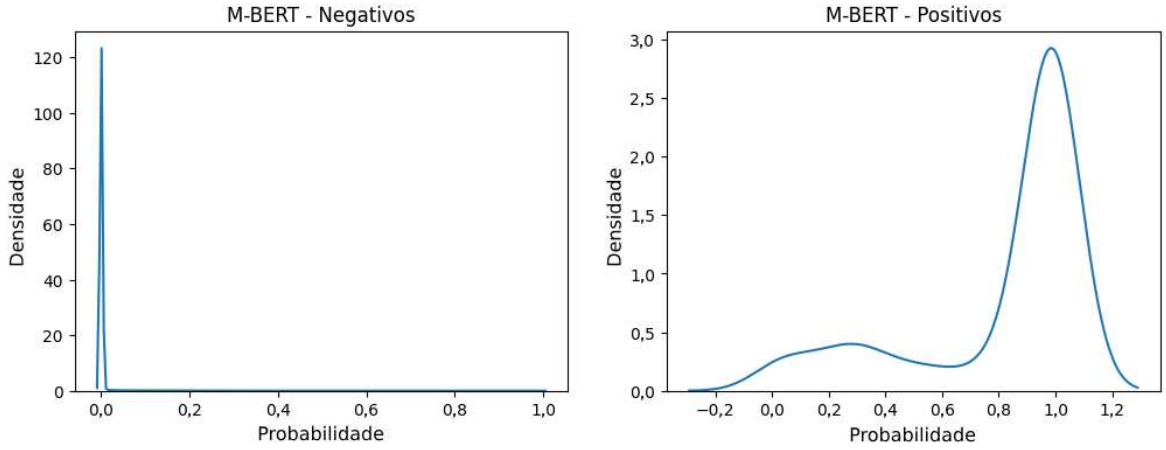


Figura C.20: Gráfico da EDK para os resultados do modelo M-BERT desbalanceado.

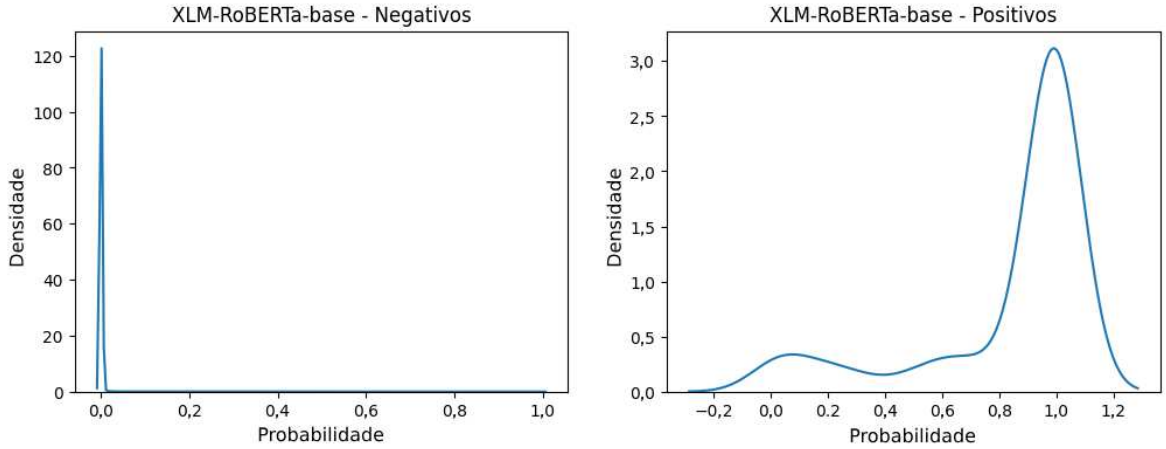


Figura C.21: Gráfico da EDK para os resultados do modelo XLM-RoBERTa-base desbalanceado.

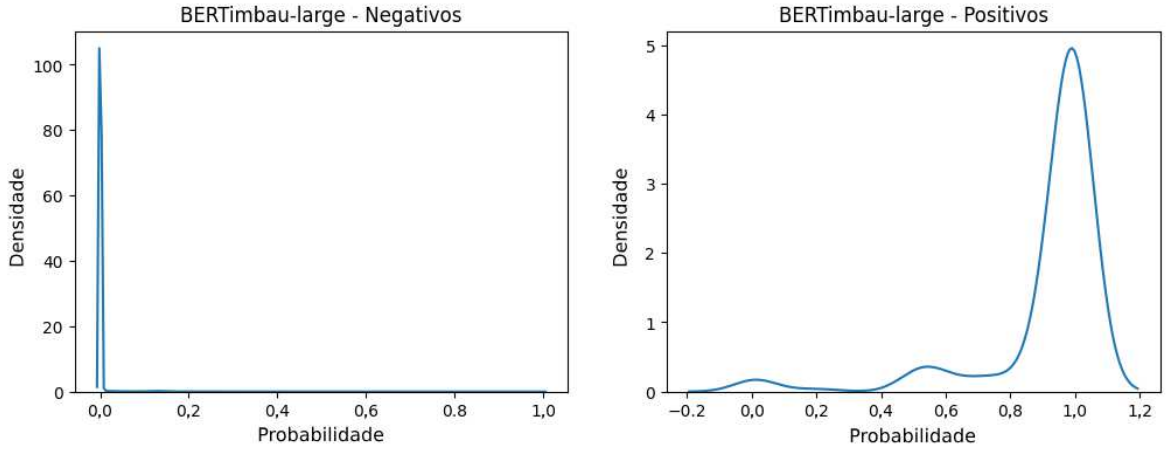


Figura C.22: Gráfico da EDK para os resultados do modelo BERTimbau-large desbalanceado.

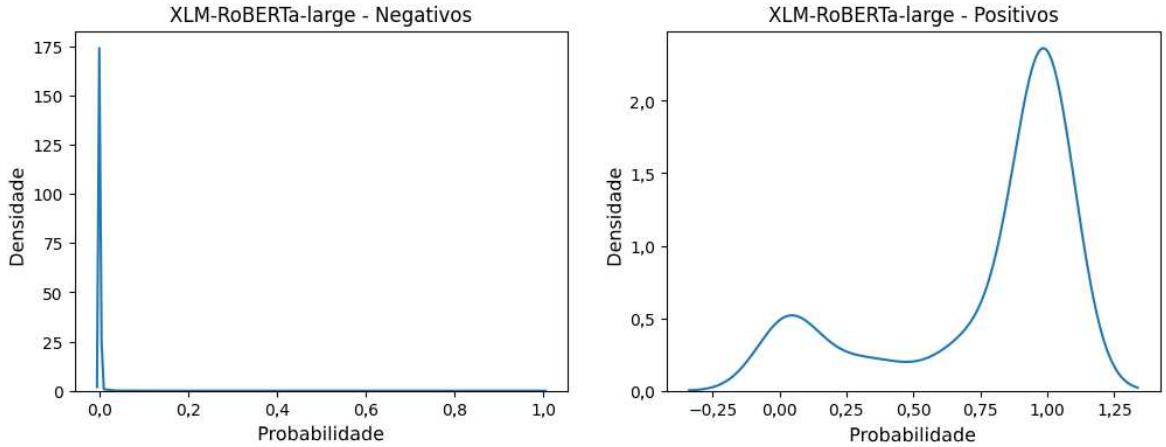


Figura C.23: Gráfico da EDK para os resultados do modelo XLM-RoBERTa-large desbalanceado.

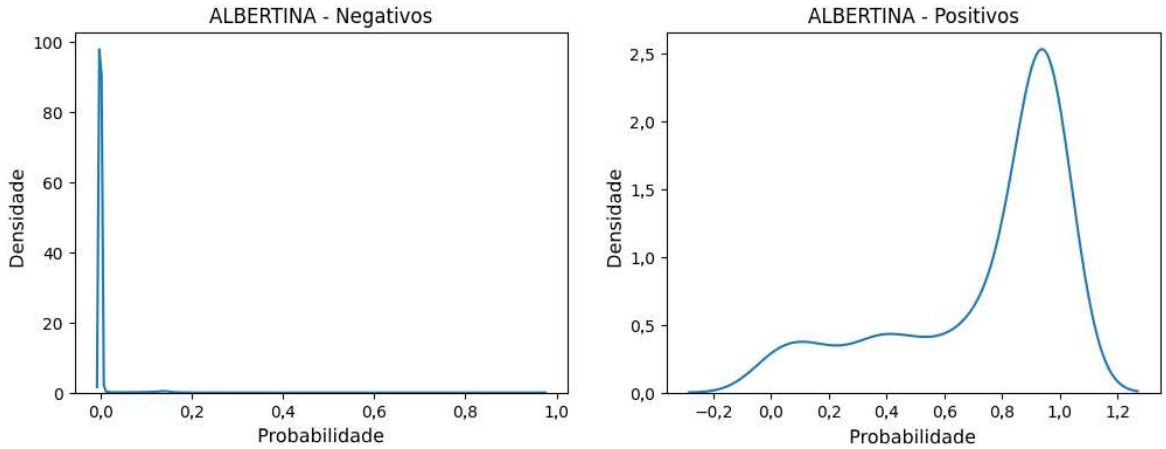


Figura C.24: Gráfico da EDK para os resultados do modelo ALBERTINA desbalanceado.