revista síntese

Desenvolvimento de uma Metodologia para a Coleta e Identificação de Atos Administrativos de Interesse nos Diários Oficiais dos Jurisdicionados do Tribunal de Contas do Estado do Rio de Janeiro (TCE-RJ)

WELLINGTON SOUZA AMARAL
Mestre em Ciência da Computação – Centro Federal de Educação Tecnológica do Rio de Janeiro.
Auditor de Controle Externo no Tribunal de Contas do Estado do Rio de Janeiro.

GUSTAVO ALEXANDRE SOUSA SANTOS
Mestre em Ciência da – Centro Federal de Educação Tecnológica do Rio de Janeiro.
Assessor no Tribunal de Contas do Estado do Rio de Janeiro.

EDUARDO BEZERRA DA SILVA
Doutor em Engenharia de Sistemas e Computação – Universidade Federal do Rio de Janeiro.
Professor titular da Escola de Informática e Computação do Centro Federal de Educação Tecnológica do Rio de Janeiro.

LEONARDO SILVA DE LIMA
Doutor em Engenharia de Produção – Universidade Federal do Rio de Janeiro.
Professor da Universidade Federal do Paraná.

AUGUSTO CÉSAR BENVENUTO DE ALMEIDA
Graduado em Engenharia da Computação – Universidade Federal de Pernambuco.
Auditor de Controle Externo no Tribunal de Contas do Estado do Rio de Janeiro.

RESUMO

O projeto “Desenvolvimento de uma Metodologia para a Coleta e Identificação de Atos Administrativos de Interesse nos Diários Oficiais dos Jurisdicionados do Tribunal de Contas do Estado do Rio de Janeiro (TCE-RJ)” aborda um desafio essencial na era digital: transformar a vasta quantidade de dados desestruturados contidos nos Diários Oficiais em informações acessíveis e úteis. Utilizando técnicas avançadas de mineração de dados, aprendizado de máquina e processamento de linguagem natural (NLP), a pesquisa visa aprimorar o controle e a fiscalização do setor público. Os Diários Oficiais, fundamentais para a transparência administrativa, frequentemente apresentam informações em formatos variados (PDF, HTML, etc.), dificultando o acesso e a análise. Este projeto propôs e testou uma metodologia inovadora baseada no modelo CRISP-DM, estruturando o processo desde a coleta de dados até a classificação de atos administrativos como nomeações e exonerações. Foram exploradas duas abordagens: o uso de *Random Forest* para dados segmentados e estruturados, e de um *Large Language Model* (LLM), como o Gemini, para analisar contextos mais complexos e textos integrais. Os resultados evidenciaram alto nível de precisão em ambas as abordagens. Enquanto o *Random Forest* destacou-se na eficiência com dados organizados, o LLM demonstrou flexibilidade ao lidar com textos variados, mantendo a acurácia mesmo em casos ambíguos.

Artigo aprovado em 29/12/2025.
DOI: https://doi.org/10.70690/m6zcsw72

Adicionalmente, a pesquisa viabilizou o uso de tecnologias emergentes, como modelos de linguagem em larga escala, para automatizar processos repetitivos, facilitando o trabalho de auditores do TCE-RJ. O estudo não apenas confirmou a viabilidade técnica da automação no setor público, mas também forneceu *insights* práticos para futuras aplicações, como a análise de contratos, licitações e convênios. Esta metodologia promete não apenas modernizar a fiscalização, mas também promover maior transparência e eficiência no controle administrativo, consolidando o uso de inteligência artificial como ferramenta estratégica na gestão pública.

Palavras-chave: Diários Oficiais; Processamento de Linguagem Natural; Random Forest; Large Language Models; Controle Externo.

Artigo aprovado em 29/12/2025.
DOI: https://doi.org/10.70690/m6zcsw72

ABSTRACT

The project "Development of a Methodology for the Collection and Identification of Administrative Acts of Interest in the Official Gazettes of Jurisdictions Overseen by the Rio de Janeiro State Court of Auditors (TCE-RJ)" addresses a critical challenge in the digital era: transforming the vast volume of unstructured data in Official Gazettes into accessible and actionable information. Using advanced techniques in data mining, machine learning, and natural language processing (NLP), the research aims to enhance control and oversight in the public sector. Official Gazettes, essential for administrative transparency, often present information in various formats (PDF, HTML, etc.), complicating access and analysis. This project proposed and tested an innovative methodology based on the CRISP-DM model, structuring the process from data collection to the classification of administrative acts such as appointments and dismissals. Two approaches were explored: the use of Random Forest for segmented and structured data, and a Large Language Model (LLM), such as Gemini, to analyze more complex contexts and full texts. The results revealed a high level of accuracy in both approaches. While Random Forest excelled in efficiency with structured data, the LLM demonstrated flexibility in handling varied texts, maintaining precision even in ambiguous cases. Additionally, the research demonstrated the feasibility of leveraging emerging technologies, such as large language models, to automate repetitive tasks, thereby facilitating the work of TCE-RJ auditors. The study not only confirmed the technical feasibility of automation in the public sector but also provided practical insights for future applications, such as analyzing contracts, tenders, and agreements. This methodology promises to not only modernize oversight but also to promote greater transparency and efficiency in administrative control, cementing the role of artificial intelligence as a strategic tool in public management.

Keywords: *Official Gazettes, Natural Language Processing (NLP), Random Forest, Large Language Models (LLMs), Public Audit.*

1 INTRODUÇÃO

A crescente digitalização de informações no setor público trouxe à tona a necessidade de metodologias eficazes para acessar, organizar e interpretar esses dados. Nos últimos anos, os Diários Oficiais passaram a ser disponibilizados em formato eletrônico, promovendo a transparência e a acessibilidade. Contudo, esses documentos ainda são amplamente subutilizados devido à falta de ferramentas robustas que possam extrair, identificar e classificar automaticamente os atos administrativos contidos nesses registros. Esse desafio é particularmente relevante para o Tribunal de Contas do Estado do Rio de Janeiro (TCE-RJ), que necessita monitorar e auditar de forma eficiente os atos administrativos de seus jurisdicionados.

Os Diários Oficiais, como instrumentos formais de publicação de atos administrativos, desempenham um papel crucial para a transparência e o controle social. Entretanto, esses documentos frequentemente contêm um volume massivo de informações não estruturadas, distribuídas em diversos formatos (PDF, HTML, entre outros), e publicadas por diferentes entes governamentais. Esse cenário complexifica a coleta e análise de dados, criando obstáculos à identificação de atos específicos, como nomeações e exonerações, que são de grande interesse para auditorias.

Atualmente, é evidente a necessidade de soluções tecnológicas que integrem técnicas avançadas de mineração de dados e processamento de linguagem natural (NLP) para transformar esses registros em fontes estruturadas de dados.

Este estudo visa desenvolver e implementar uma metodologia para a coleta, identificação e classificação automática de atos administrativos publicados nos Diários Oficiais. Os objetivos específicos incluem:

Mapear fontes de publicação: Identificar as fontes e formatos dos Diários Oficiais dos jurisdicionados do TCE-RJ.

Propor uma metodologia robusta: Estruturar um processo baseado no modelo CRISP-DM, incluindo etapas de coleta, processamento e análise de dados.

Desenvolver ferramentas tecnológicas: Implementar algoritmos de aprendizado de máquina, como *Random Forest* e *Large Language Models* (LLMs), para extrair e classificar atos administrativos de interesse.

Validar a metodologia proposta.

Promover a escalabilidade: Estabelecer um protótipo funcional que possa ser expandido para outros órgãos e tipos de publicações administrativas.

2 DESENVOLVIMENTO

2.1 Referencial Teórico

O primeiro passo em busca de alcançar os objetivos da pesquisa é a busca por trabalhos correlatos com a finalidade de identificar como o problema tem sido abordado na literatura nacional e internacional. Nesse sentido, a presente seção se propõe a apresentar a revisão bibliográfica dos trabalhos acadêmicos que atenderam às questões de pesquisa correlatas à temática deste estudo. A temática abordada continha dois grandes temas: 1) extração de informações em diários oficiais: utilizando as *strings* de busca (*Information, extraction, official gazette*); e 2) métodos de classificação de dados vindos de diários oficiais: utilizando as *strings* de busca *"text classification" and "official gazette"*.

Ambos os temas foram buscados nas bases *Web of Science*, *Scopus* e *Google Scholar*.

2.1.1 Extração de informações em diários oficiais

Diversos trabalhos abordam a extração de informações de documentos jurídicos. Em Constantino *et al.* (2023), os autores propõem uma abordagem independente de estrutura aplicável a vários documentos governamentais, incluindo diários oficiais, e aproveita a aprendizagem ativa para **minimizar** os esforços de rotulagem manual para melhorar a eficiência.

A ferramenta DODFMiner (Guimaraes *et al.*, 2024) oferece uma abordagem em três etapas para extração de informações de diários oficiais. Essa abordagem foi projetada especificamente para o Diário Oficial do Distrito Federal.

Xavier *et al.* (2015) utilizaram a técnica de indexação de dados textuais por meio da medida TF/IDF, a fim de processar documentos publicados no Diário Oficial do Município de Cachoeiro de Itapemirim-ES. Por meio das etapas de pré-

processamento de texto, os autores construíram uma máquina de busca específica para esses documentos para facilitar a tarefa de recuperação deles. Para fins de avaliação, os autores rotularam manualmente 198 entradas em diferentes categorias. Em seguida, a máquina de busca proposta foi usada para recuperar os documentos.

Em Pinto *et al.* (2021), os autores utilizaram técnicas de expressões regulares para a extração de dados de diários oficiais dos estados do Rio de Janeiro, Maceió, Palmas, Recife e Florianópolis. O escopo do artigo foi limitado aos atos públicos envolvendo movimento de pessoal (publicações envolvendo nomeações, demissões, e nomeações para cargos comissionados) num período de nove anos. Os dados extraídos foram representados num *Resource Description Framework* (RDF)/XML.

Em Ogawa *et al.* (2016) os autores extraíram informações dos Diários Oficiais Japoneses e dos Diários Oficiais Japoneses Versão Inglês no período de 1942 a 1956 com a finalidade de aumentar o dicionário bilíngue de termos legais que atualmente contém somente 3872 entradas. Com as informações extraídas de ambos os diários oficiais, as ferramentas GIZAM++ e Moses (*Open source toolkit for statistical machine translation*) foram utilizadas para ampliar o dicionário bilíngue com mais 820 frases.

2.1.2 Métodos de classificação de dados vindos de diários oficiais

Vários estudos exploraram o uso de técnicas de classificação de texto para detecção de fraudes em vários domínios. Em Luz de Araujo *et al.* (2020), os autores criaram um conjunto de dados de documentos do Diário Oficial rotulados como fraudes e irregularidades. Em seguida, realizaram uma comparação de modelos tradicionais de aprendizado de máquina com uma abordagem de aprendizagem por transferência (*transfer learning*) baseada em ULMFiT para esta tarefa.

2.1.3 Métodos híbridos

Em Neves Junior *et al.* (2018), os autores propõem um método para a extração de informações do Diário Oficial do Estado de Pernambuco e implementaram um

motor de inferência baseado em árvore de decisão (*software Weka*) para a predição de resultados de sindicâncias.

Em Berrazega *et al.* (2016a; 2016b) propõem uma abordagem automática baseada em conhecimento para identificar e anotar (semanticamente) as categorias de textos normativos árabes coletados do Diário Oficial da República da Tunísia. Essa abordagem combina (1) uma taxonomia de categorias de disposições normativas, (2) uma base terminológica normativa e (3) um anotador semântico baseado em regras. Para construir o anotador semântico (detalhado em Berrazega *et al.*, 2016c), os autores definiram manualmente um conjunto de regras de anotação da forma "*if*" condição então ação. Como resultado, os autores produziram uma base de 35 regras que, ao todo, cobrem 14 categorias normativas. O desempenho da abordagem foi avaliado em termos de precisão, *recall* e pontuação F para categorizar instâncias em 14 categorias normativas. Os resultados obtidos no conjunto de dados de teste foram 96,4% para Precisão, 96,06% para *Recall* e 96,23% para F-score.

Em Rocha (2011), o autor faz a obtenção de dados do Diário Oficial da União com as publicações das ações do Governo Federal do Brasil. O estudo visa analisar os dados obtidos para identificação de irregularidades. Foram extraídos os dados com as informações de "Contratado" e "Contratante" e foi utilizada a ferramenta comercial ASG-CYPRESS instalada na Controladoria Geral da União (CGU). A metodologia CRISP-DM foi utilizada para o entendimento dos dados.

Constantino *et al.* (2022) abordam dois aspectos fundamentais no processamento de Diários Oficiais: a segmentação de trechos textuais e a classificação semântica desses trechos. Para a segmentação, os autores propuseram uma heurística orientada à estrutura dos documentos, capaz de identificar e extrair blocos de texto relevantes com base em características espaciais e visuais. Já para a classificação semântica, foi desenvolvida uma estratégia baseada em aprendizado de máquina ativo, utilizando classificadores transformers de última geração. Essa abordagem permite reduzir significativamente o esforço manual de rotulagem, ao selecionar automaticamente os trechos mais informativos para anotação. Como resultado, foi implementado um protótipo de ferramenta de anotação integrada a um serviço Web, que se conecta diretamente ao processo de classificação. Os experimentos realizados indicaram uma acurácia de 85% na classificação semântica

e um valor de MacroF1 de 75%, demonstrando a robustez e eficiência da solução proposta.

Em Rodríguez e Bezerra (2020), os autores utilizaram processamento de linguagem natural para automatizar o reconhecimento de Entidades Nomeadas (Agentes Públicos) em uma base de Portarias e se preocuparam com os atos de Nomeação e Exoneração. E por meio das técnicas aplicadas de tokenização, *pos-tagger* e *chunk* foi demonstrada a análise, classificação e marcação semântica para cada sentença e palavra, numa dada linguagem, atribuindo classificações como substantivo, verbo, entre outros, possibilitando a extração dos trechos contendo Nomes Próprios, objetivo principal do artigo.

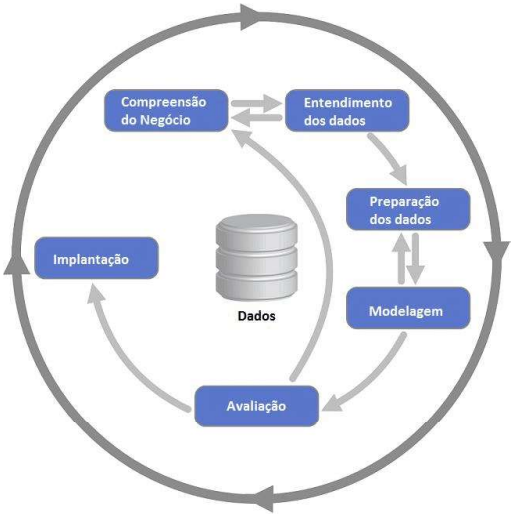
Conforme OKBR (2024), a iniciativa do Querido Diário, ao tratar da coleta, processamento e disponibilização de dados públicos municipais, alinha-se aos objetivos desta pesquisa. O projeto Querido Diário é uma iniciativa de código aberto que coleta e organiza dados dos diários oficiais municipais do Brasil, utilizando a linguagem *Python* e *frameworks* como *Scrapy* para raspagem de dados, Pandas para manipulação e análise, e *SpaCy* para processamento de linguagem natural. Por meio dessas tecnologias, o projeto extrai e estrutura informações de atos administrativos, contratos e editais, tornando-os acessíveis e pesquisáveis. A arquitetura modular do projeto permite a contribuição de desenvolvedores, visando a ampliação da cobertura dos diários oficiais e a melhoria contínua da ferramenta.

2.2 Metodologia

A metodologia adota o modelo *CRoss Industry Standard Process for Data Mining* (CRISP-DM), adaptado para atender aos objetivos específicos deste projeto de pesquisa.

CRISP-DM é o modelo reconhecido como o padrão mundial da indústria para os processos de mineração de dados. Ele descreve as abordagens comumente utilizadas por especialistas em mineração de dados para a solução de problemas, independentemente da área de negócio e das tecnologias aplicadas. (Shearer, 2000). A Figura 1 ilustra essas etapas.

Figura 1 – Etapas do processo CRISP-DM.



Fonte: Shearer (2000).

O CRISP-DM é um modelo de processo iterativo e cíclico, o que significa que as fases não são rígidas ou sequenciais. Em vez disso, elas são interativas e podem ser revisitadas conforme necessárias ao longo do projeto. Embora o CRISP-DM apresente uma sequência lógica de fases, na prática, o fluxo entre as fases é flexível e adaptável às necessidades específicas do projeto. Isso permite uma abordagem mais dinâmica e responsiva aos objetivos desse projeto de pesquisa (Shearer, 2000).

A fase de Compreensão do Negócio foca em entender os elementos de negócio que se relacionam com o objetivo de desenvolver uma metodologia para coletar e identificar atos jurídicos de interesse nos DOs dos jurisdicionados do TCE-RJ. Nessa fase está inserido o objetivo específico 1 Realizar Revisão Bibliográfica, cujo propósito é pesquisar trabalhos correlatos que abordam a coleta e classificação de atos oficiais, identificando as melhores práticas, algoritmos e técnicas utilizadas.

A fase de Entendimento dos Dados foca no reconhecimento de dados e início de atividades para familiarização com os dados, identificando problemas ou conjuntos interessantes. Nessa fase, são identificadas as fontes de dados e formatos disponíveis, bem como os diversos formatos de publicação utilizados por diferentes órgãos. Isso envolve a compreensão da diversidade de dados a serem tratados, onde está centrada o objetivo específico 2 Identificar Fontes e Formatos. Nessa etapa, a atividade de compreensão dos dados é fundamental para o desenvolvimento da metodologia proposta. Dessa forma, nosso objetivo é obter um entendimento maior dos dados disponíveis nos Diários Oficiais dos jurisdicionados do TCE-RJ.

Para o desenvolvimento eficaz da metodologia de coleta e identificação de atos jurídicos nos Diários Oficiais, é fundamental analisar várias características da divulgação do Diário Oficial. Isso inclui o formato do arquivo (como PDF, HTML, imagem etc.), que pode afetar a estratégia de extração de dados. A abrangência dos órgãos jurisdicionados é outro aspecto importante, pois determina a amplitude dos dados a serem coletados. A existência de uma interface WEB é crucial, pois pode facilitar o acesso e a coleta de dados. Além disso, se o documento é pesquisável por texto, o que tende a simplificar o processo de identificação de atos jurídicos específicos. Cada uma dessas características desempenha um papel importante na definição da abordagem de coleta de dados e na eficácia da metodologia proposta.

A fase de Preparação dos Dados foca na construção do conjunto de dados final a partir dos dados iniciais. Nessa pesquisa, ela ocorrerá em várias vezes no processo. No caso específico em tela, essa fase ocorrerá com mais frequência para atingir os objetivos específicos 3. Desenvolver Metodologia.

A fase de Modelagem, para este projeto de pesquisa envolve realizar um estudo abrangente dos algoritmos de NLP disponíveis, selecionar os algoritmos mais adequados para a tarefa de identificação de atos jurídicos nos textos dos DO, adaptá-los à metodologia geral e realizar testes de desempenho para avaliar sua eficácia e precisão para identificar atos de nomeação e exoneração nos Diários Oficiais. Nessa fase, os modelos são treinados e ajustados com os dados preparados na fase anterior. Além disso, técnicas de classificação são utilizadas para categorizar os atos de acordo com suas características. É importante notar que a modelagem é um processo iterativo, o que significa que os modelos podem ser continuamente refinados e ajustados com base nos resultados obtidos e nas necessidades do projeto. Essa fase é crucial para a eficácia da metodologia, pois é onde os algoritmos começam a aprender e a categorizar com base nos dados disponíveis.

A fase de Avaliação no CRISP-DM para este projeto de pesquisa é fundamental para garantir que os modelos desenvolvidos na fase de modelagem estão funcionando conforme o esperado. São realizados testes e avaliações para medir eficácia, precisão e capacidade de classificação do modelo. Os resultados obtidos são compilados e as conclusões são devidamente documentadas. Os resultados obtidos são compilados e documentadas as conclusões alcançadas.

Na fase de Implantação do CRISP-DM para este projeto de pesquisa, o protótipo funcional de coleta e identificação de atos jurídicos será desenvolvido e colocado em uso prático a fim de testar a metodologia proposta. Nessa fase está inserido o objetivo “Desenvolver Protótipo”. Esse protótipo fará a obtenção dos dados baseando-se nos modelos de aprendizado de máquina e NLP que foram treinados e ajustados nas fases anteriores. O protótipo será integrado aos Diários Oficiais identificados, permitindo a aplicação prática da metodologia. Além disso, os resultados obtidos pelo protótipo serão apresentados de forma acessível aos *stakeholders*, garantindo que os conhecimentos descobertos sejam utilizados efetivamente na tomada de decisão. Essa fase também pode envolver o monitoramento contínuo do desempenho do protótipo e a realização de ajustes conforme necessário para garantir que a solução continue a atender às necessidades do projeto. A implantação do protótipo é uma fase crítica que garante que o trabalho realizado nas fases anteriores se traduza em valor prático.

Cada uma dessas fases é crucial para o desenvolvimento de uma metodologia eficaz para a coleta e identificação de atos jurídicos de interesse nos DOs dos jurisdicionados do TCE-RJ.

2.3 Resultados da pesquisa

A partir desta seção, os resultados da pesquisa são apresentados. Os resultados obtidos foram organizados de acordo com as etapas previstas pela metodologia proposta.

2.3.1 Entendimento dos dados

Nesta fase, realizamos uma análise abrangente dos Diários Oficiais dos jurisdicionados do TCE-RJ. Identificamos as fontes de dados, analisamos os formatos dos Diários Oficiais.

Os resultados desta etapa fornecem uma base sólida para o desenvolvimento das etapas subsequentes da metodologia, incluindo a preparação dos dados, a

modelagem e a avaliação dos modelos. Eles também nos permitem refinar nossa abordagem e garantir que estamos focados nos dados mais relevantes e úteis para a coleta e identificação de atos jurídicos de interesse.

Nessa fase, foram identificadas 94 fontes diferentes de entes e órgãos públicos que geram informação para os diários oficiais. A partir dos diários oficiais identificados, foi realizada uma análise de seus web sites e interfaces com vistas a identificar as seguintes características: a URL, qual entidade faz a veiculação de matérias naquele veículo, a qual (Executivo, Legislativo ou Judiciário) poder pertence, se o diário oficial eletrônico tem um interface web própria para acessar as matérias, se há sistema de proteção contra coleta automatizada, como CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*), qual o formato do arquivo do Diário Oficial, se é possível descarregar o arquivo inteiro do diário de um dia ou se é necessário descarregar vários arquivos para compor um único dia de publicação e se é possível realizar pesquisa e extração do texto no arquivo disponibilizado.

Dentre as 94 fontes de informações de Diários Oficiais, foram analisadas 28, o que representou 30% do total, conforme pode ser visto na **Tabela 6**. Em termos de materialidade, esses 28 entes foram responsáveis por 84% do valor total empenhado por todos os jurisdicionados do TCE-RJ no período de 2017 a 2022, resultando nos dados descritos na **Tabela 7**.

Tabela 6 – Diários oficiais (DOs) identificados e analisados.

Avaliação Quantitativa dos DOs	Quantidade
DOs identificados	94
DOs analisados	28
DOs analisados (%)	30%

Fonte: Elaborado pelos autores.

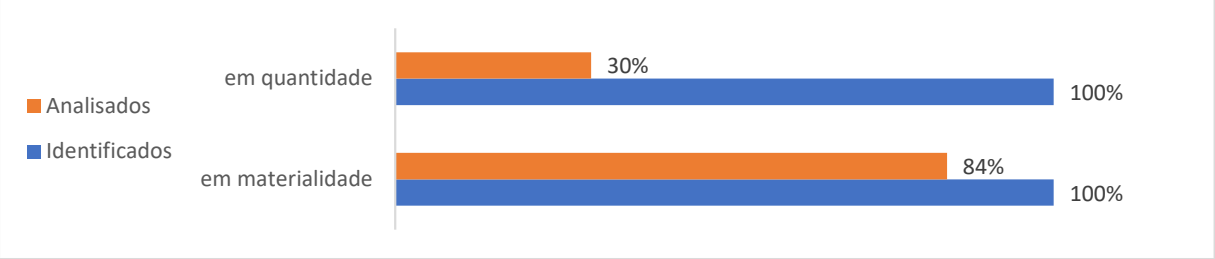
Tabela 7 - Materialidade dos DOs.

Avaliação Quantitativa dos DOs	Recursos (em Milhões R$) *
DOs identificados por materialidade	R$ 303.540
DOs analisados por materialidade	R$ 254.346
DOs analisados por materialidade (%)	84%

*Recursos empenhados entre 2017 e 2022 dos jurisdicionados do TCE-RJ.

Fonte: Elaborado pelos autores.

Figura 2 – Gráfico do percentual dos diários oficiais analisados comparando-se a materialidade e a quantidade.



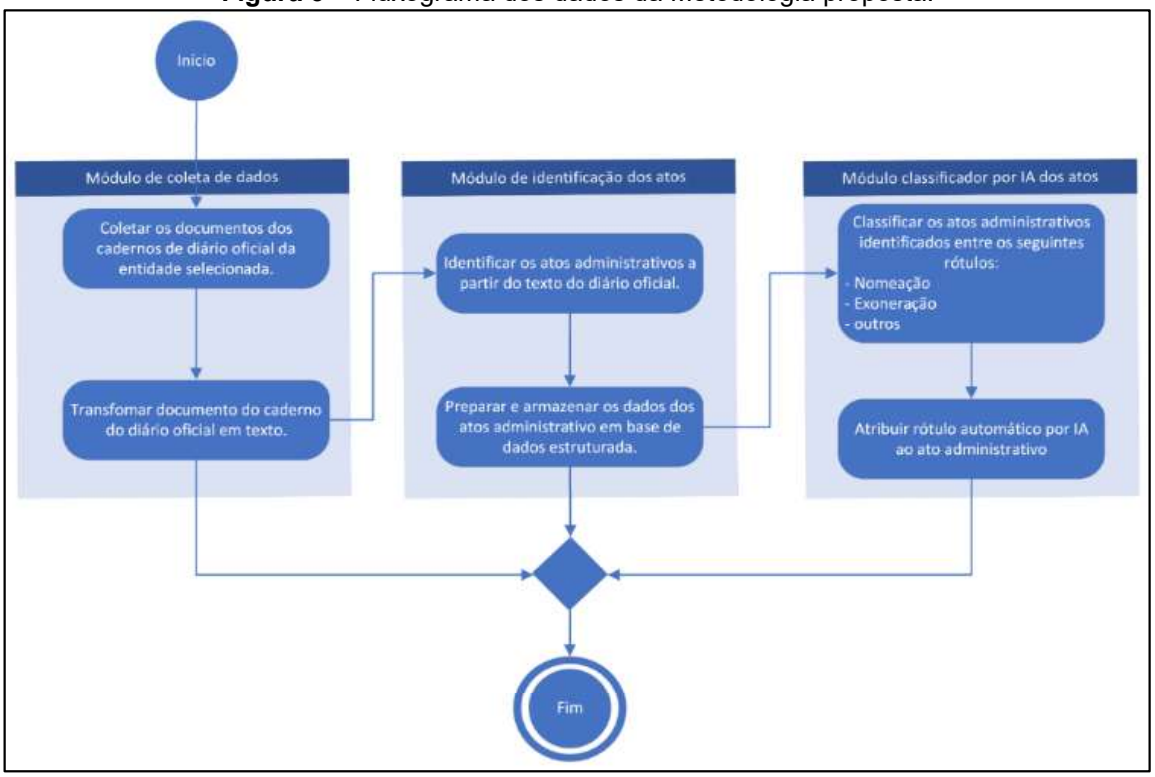
Fonte: Elaborado pelos autores.

A análise pela materialidade ajuda a determinar a relevância e o significado de certos dados ou informações em relação aos objetivos gerais da pesquisa. As tabelas e gráficos a seguir cotejam essas informações.

2.3.2 Preparação dos dados

Nesta etapa, foi construída uma metodologia para extrair, identificar e classificar os atos administrativos dos diários oficiais dos jurisdicionados do TCE-RJ, conforme apresentada a **Figura 3**.

Figura 3 – Fluxograma dos dados da Metodologia proposta.



Fonte: Elaborado pelos autores.

A seguir, cada um dos módulos apresentados pela metodologia proposta será descrito.

a) Módulo de Coleta de Dados

No módulo de coleta de dados, o processo inicia com a coleta dos documentos dos cadernos de diário oficial na fonte de informação da entidade selecionada. As fontes utilizadas para obtenção dos cadernos podem incluir o desenvolvimento de robôs raspadores automatizados (*scrapers*), que consultam os portais e capturam os cadernos, a coleta manual ou o uso de soluções terceirizadas que indexam os diários oficiais de entes públicos, como o projeto Querido Diário da *Open Knowledge* Brasil.

Esta etapa é fundamental, pois envolve a obtenção das informações publicadas oficialmente que serão analisadas e processadas nas etapas subsequentes.

Durante a coleta são registradas informações acerca do processo, qual sejam, por exemplo: entidade selecionada, data de coleta, data do caderno e nome do arquivo do documento.

Após a coleta, o documento do caderno do diário oficial é transformado em texto puro. Uma vez transformados, esses dados ficam disponíveis para pesquisa textual por ferramentas de busca, como o Bing, por exemplo. Esse formato acessível permite que usuários realizem buscas específicas dentro do texto do diário oficial, facilitando a localização de informações pertinentes.

b) Módulo de Identificação dos Atos

O Módulo de Identificação dos Atos foi desenvolvido com o objetivo principal de identificar e segmentar as matérias publicadas nos Diários Oficiais em atos administrativos. Após essa identificação inicial, os dados são organizados e armazenados em uma base de dados estruturada, permitindo que o conteúdo dos Diários Oficiais seja facilmente acessível e pesquisável. Esse armazenamento segmentado possibilita aos usuários realizarem buscas específicas de acordo com suas necessidades, promovendo maior agilidade na localização de informações relevantes.

c) Módulo Classificador por Inteligência Artificial dos Atos

Já o Módulo Classificador por Inteligência Artificial (IA) dos Atos aplica algoritmos de IA incluindo um modelo de linguagem de larga escala (LLM), para classificar os atos administrativos identificados. Os atos são rotulados automaticamente em categorias pré-definidas: Nomeação, Exoneração. O uso do LLM expande a compreensão do contexto linguístico dos textos, melhorando a precisão na atribuição dos rótulos, especialmente em casos complexos onde termos como "nomeação" e "exoneração" podem aparecer em um mesmo parágrafo. Esses rótulos enriquecem os dados, fornecendo informações adicionais sobre a natureza de cada ato administrativo, o que facilita consultas detalhadas e específicas por parte dos usuários.

2.4 Modelagem

Foram desenvolvidas duas abordagens distintas ao longo do processo de modelagem e refinamento. A primeira abordagem utilizou o modelo *Random Forest* aplicado à amostra A1, que consiste em 3.501 parágrafos extraídos com expressões regulares, representando textos que foram classificados como atos de nomeação ou exoneração. Já a segunda abordagem foi implementada utilizando um *Large Language Model* (LLM) para lidar com a amostra A2, composta por arquivos PDF de 105 edições de Diários Oficiais do Estado do Rio de Janeiro, e dos municípios Angra Dos Reis, Armação Dos Búzios, Duas Barras, Engenheiro Paulo De Frontin, Mendes, Vassouras, Aperibé, Areal, Consórcio Intermunicipal De Saúde Da Região Serrana, Associação Municípios, Belford Roxo, Campos Goytacazes, Casimiro de Abreu, Cordeiro, Iguaba Grande, Macaé, Maricá, Mesquita, Miguel, Pereira, Niterói, Nova Iguaçu, Quatis, Queimados, Quiçamã, São Joao De Meriti, São Jose Do Vale Do Rio Preto e São Pedro Da Aldeia, cobrindo publicações entre 2 de agosto e 1 de outubro de 2024. Ambas as abordagens evoluíram de forma iterativa nas fases de modelagem e avaliação do CRISP-DM, resultando em abordagens distintas para classificação de atos administrativos. As sessões a seguir apresentam as ações e resultados segregadas por abordagem.

2.4.1 Abordagem classificador Random Forest

Para a abordagem pelo classificador *Random Forest*, foram selecionados 3.501 parágrafos por meio de expressão regular, os quais representaram textos para serem classificados, se eram pertencentes de nomeação ou exoneração. A amostra foi segregada 80% para o conjunto de treino e 20% para o conjunto de teste. Para essa etapa, foram utilizadas em conjunto a técnica de Tokenização, e o método de classificação *Random Forest*.

A tokenização é uma das primeiras etapas de preparação durante o desenvolvimento de NLP e tem a finalidade de fracionar o texto de entrada em partes menores na forma de subconjuntos de caracteres, chamados de tokens. Esses tokens são utilizados em várias tarefas de NLP, como análise morfológica, marcação de classe de palavras e análise, que são tratamentos subsequentes para contagem de tokens, identificação de radicais de palavras, verificação sintática, entre outras (Grefenstette, 1999).

Sob a mesma finalidade, o classificador proposto neste estudo precisa ter essa etapa de pré-processamento das sentenças extraídas da amostra inicial dos DO's do Estado do Rio de Janeiro (DOERJ). Para esse processo de tokenização, foi utilizado o tokenizador DistilBERT (Sanh, 2019), disponível pela biblioteca HuggingFace, bastante difundido na comunidade NLP para tarefas de tokenização. O DistilBERT, sendo uma versão minimalista do BERT (Devlin *et al.*, 2019), herda essa estratégia de tokenização chamada "WordPiece".

O WordPiece é um método de tokenização subpalavra que divide palavras em unidades menores, chamadas de *subwords*. Isso permite que o modelo lide com palavras raras ou fora do vocabulário, quebrando-as em *subwords* mais comuns. Sendo assim, a principal diferença entre DistilBERT e BERT refere-se ao tamanho do vocabulário para reduzir o tamanho do modelo e aumentar a eficiência computacional.

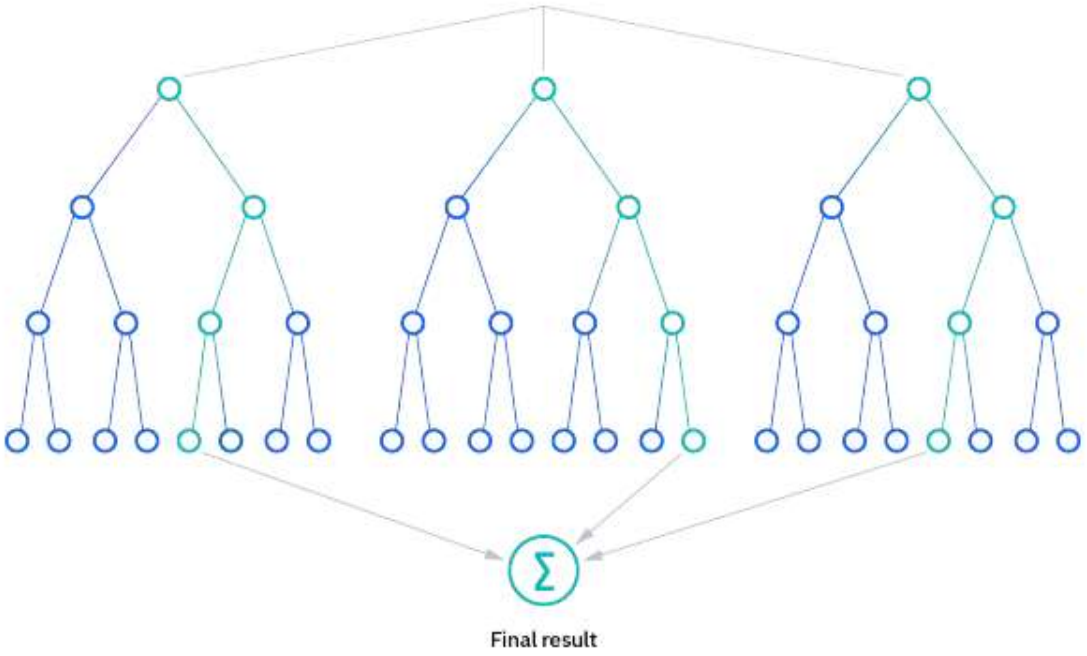
As vantagens desse processo de tokenização é a possibilidade de obter um melhor equilíbrio entre o tamanho do vocabulário e a capacidade de representar palavras raras ou desconhecidas. Além disso, há também uma identificação para tokens especiais do tipo: ínicio das frases (CLS), separação de frases (SEP) e preenchimento (PAD) específico. Em resumo, o tokenizador DistilBERT, baseado no

WordPiece, é fundamental para preparar o texto para o modelo. Ele divide o texto em unidades menores, permitindo o processamento de palavras dentro e fora do vocabulário, além de fornecer informações estruturais importantes para o modelo.

O algoritmo *Random Forest* (RF) é uma técnica de perturbação e combinação projetada para Árvores de Decisão (AD) (Breiman *et al.*, 1984), o qual representa um conjunto diversificado de classificadores de AD criados sob critérios de aleatoriedade. Ao final do procedimento, a previsão do conjunto é dada como a previsão média dos classificadores individuais.

A técnica de RF foi introduzida por Breiman (2001), aplicando amostragem de bootstrap e seleção aleatória de atributos para criar um conjunto de AD não correlacionadas, representando uma floresta aleatória. Nessa etapa, cada árvore é treinada em uma amostra diferente dos dados e considera-se apenas um subconjunto aleatório de recursos em cada divisão. Essa aleatoriedade induzida reduz a variância do modelo e melhora a generalização. A previsão final é obtida pela agregação das previsões de todas as árvores, geralmente por meio de voto majoritário (classificação), conforme a **Figura 4** apresenta a seguir.

Figura 4 – Execução do Random Forest.



Fonte: Kavlakoglu (2024).

Com o conjunto de AD (Breiman et al., 1984) gerado na execução do RF, como a **Figura 4** mostra. O RF pode ser caracterizado por três principais parâmetros (ou hiper parâmetros): tamanho do nó, número de árvores e número de atributos

(Breiman, 2001). Esses hiper parâmetros tem a função de otimizar o desempenho do modelo nas tarefas, podendo ser classificação ou regressão. Durante a execução do RF, cada AD é treinada em uma amostra inicial, a qual é aleatória e com reposição do conjunto de dados original. Uma outra parte dos dados, conhecida como amostra out-of-bag (OOB), é reservada para validação (Breiman, 2001).

A predição final do RF é determinada pela agregação das previsões de cada árvore individual. Em problemas de regressão, a previsão final é a média das previsões de cada árvore, enquanto em problemas de classificação, a classe mais frequente, determinada por voto majoritário, é a previsão final (Breiman, 2001).

Nessa etapa, foram realizadas as tarefas de classificação e agrupamento pela implementação das rotinas computacionais em linguagem de programação Python. Foram utilizadas as bibliotecas do pacote *transformers*, que oferece funções para aplicação da técnica de tokenização, e o pacote *scikit-learn* (Pedregosa, 2011), que possui funções para a aplicação de diversas técnicas de balanceamento em conjuntos de dados e permite o treinamento do classificador *Random Forest*.

2.4.2 Abordagem classificador LLM

Para explorar uma abordagem alternativa ao modelo *Random Forest* na classificação de atos administrativos extraídos dos Diários Oficiais, foi desenvolvido um classificador baseado em um *Large Language Model* (LLM). Essa abordagem visa avaliar o desempenho do modelo LLM em comparação com o *Random Forest*, especialmente em relação à sua capacidade de lidar com o contexto linguístico dos documentos oficiais.

2.5 Avaliação

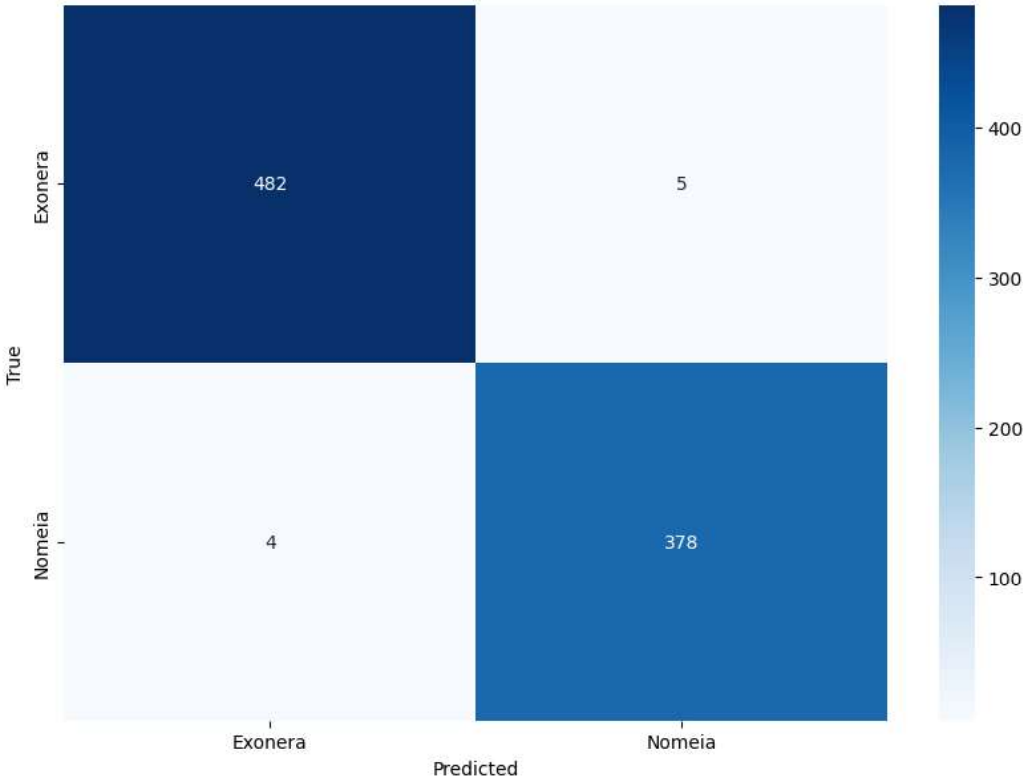
Esta seção apresenta uma análise dos resultados coletados em todo o estudo, com o intuito de identificar observações relevantes que embasarão as etapas subsequentes da pesquisa. Os resultados foram explorados de maneira expositiva e descritiva, permitindo uma compreensão ampla dos principais achados e das

características observadas nos dados quando comparadas as abordagens de classificação usando 1) *Randon Forest* e 2) LLM, conforme descrito nas seções a seguir.

2.5.1 Abordagem Classificador Random Forest

Aplicando-se as técnicas descritas nas etapas anteriores, o classificador desenvolvido classificou corretamente 99% das amostras do conjunto de teste. A matriz de confusão identifica quantos exemplos de cada classe foram corretamente classificados e quantos foram erroneamente classificados em outras classes. Na **Figura 5**, a visualização em formato de mapa de calor facilita a interpretação dos resultados.

Figura 5 – Matriz de Confusão dos resultados do classificador *Random Forest*.



Fonte: Elaboração pelos autores.

2.5.2 Abordagem Classificador LLM

Este procedimento de classificação consiste na utilização do LLM do Google, o Gemini Flash - versão 1.5, mais informações podem ser encontradas na página oficial: https://ai.google.dev/gemini-api/docs). Nesse sentido, utilizou-se do modelo disponível em modo gratuito (*free tier*), a fim de não inviabilizar a pesquisa durante a análise de conteúdo a partir do texto extraído dos PDFs.

Sendo assim, cada texto de PDF é executado pelo LLM em dois prompts específicos: um para identificar nomeações e outro para exonerações. Esses prompts foram elaborados para guiar o modelo a extrair as informações relevantes. Na **Figura 6**, é apresentado para fins de exemplificação o *prompt* usado para identificar atos de nomeação e estruturar as informações relevantes em formato JSON.

Figura 6 – *Prompt* usado para identificar atos de nomeação e estruturar as informações relevantes em formato JSON.

```
prompt_nomeacao = """Identifique e extraia todos os atos de nomeação de servidores públicos dos documentos apresentados. \
Um ato de nomeação refere-se ao procedimento administrativo em que um indivíduo é formalmente designado para ocupar um cargo público ou administrativo.
Identifique os atos de nomeação no texto quando houver, senão retorne [].

Critérios para identificar os atos de nomeação:

Frases que começam com "Nomear [nome]" seguido pela descrição do cargo ou função para o qual a pessoa está sendo nomeada.
O ato deve envolver o preenchimento de um cargo público, como cargos comissionados (CC) ou funções gratificadas (FG).
Desconsidere atos que exoneram, dispensam ou designam alguém de um cargo. Esses não são atos de nomeação.

Exemplos de atos de nomeação (corretos):
"Nomear Anderson Guedes Ribeiro, para ocupar o cargo de provimento em comissão de Assessor de Secretaria, símbolo CC1, da(o) Instituto de Educação."
"Nomear Jackeline de Fátima Pena, para ocupar o cargo de provimento em comissão de Chefe de Setor, símbolo CC4, da(o) Secretaria Municipal Saúde."
"Nomear Norma Suely Cristina Cataldo Izoldi para exercer a função gratificada de Assistente Pedagógico de Empreendedorismo, símbolo FGPE."

Exemplos de atos que não são nomeação (devem ser ignorados):
"Exonerar Fábio Castilho de Souza, da função gratificada de Assistente Operacional, símbolo FGS, da(o) Secretaria Municipal de Educação."
"Exonerar Plínio Marcus Dutra Pinheiro, do cargo de Gerente de Atendimento, símbolo CC3, da(o) Instituto de Educação do Município de Resende."
"Designar o servidor Valmer Feres Vignoli, matrícula 46496, como Supervisor de Proteção Radiológica do Hospital Municipal de Meriti."
"Designados os servidores Fernando Cruz; como Titular e Nilton Campos, como Suplente, para atuar na Unidade de Controle Interno Setorial."
"Dispensar os membros titulares e suplentes ao Conselho Deliberativo Municipal de Trabalho Emprego e Renda nos termos do Decreto Municipal."
"CONCEDER ao funcionário MIGUEL RAIMUNDO PAES, Fiscal de Tributos Municipal, Nível 7E, Matrícula nº3457 da Secretaria Municipal de Fazenda Licença Prêmio pelo prazo de 02(dois) meses."

Para cada ato de nomeação identificado (apenas se houver), extraia as seguintes informações:

filename: nome do arquivo processado
nome: nome da pessoa nomeada.
matricula: matrícula da pessoa nomeada do cargo.
data: data de publicação da nomeação identificada no documento em formato DD-MM-YYYY.
cargo: cargo ou função para o qual a pessoa está sendo nomeada.
ato: Nomeação.
paragrafo: parágrafo completo do documento onde o ato de nomeação foi mencionado.
pagina: número de página do documento onde o ato foi identificado.
unidade: unidade gestora onde a pessoa foi nomeada.
municipio: município onde a pessoa foi nomeada.

Apresente os resultados extraídos como um arquivo JSON, com um objeto para cada nomeação identificada. Cada objeto deve conter os campos listados acima."""
```

Fonte: Elaboração pelos autores.

O modelo Gemini foi configurado com uma “temperatura=0.1”, “top_k=3" e “top_p=0.3” a fim de mitigar comportamentos flexíveis de geração de texto, esse tipo de comportamento característico de modelos LLM tende a promover um problema, popularmente conhecido como “alucinação” ou, em inglês, “*hallucination*” (JI et al., 2023) que vide a capacidade de criação de texto desses modelos, pode resultar em conteúdo equivocado. A escolha dessas configurações visa limitar a criatividade do modelo com a precisão das respostas. Dessa forma, executa-se cada um dos prompts

para extrair os atributos especificados e estruturá-los em resposta com formato JSON a cada prompt.

A **Figura 7** apresenta o resultado estruturado em formato JSON, no qual são apresentadas informações detalhadas sobre exonerações de servidores públicos que, embora convocados, não assumiram seus cargos dentro do prazo estipulado pela legislação. Cada entrada do JSON inclui campos específicos, como "cargo", "nome", "data", "página", "parágrafo" e o "tipo" de ato (neste caso, "Exoneração"). Esses detalhes facilitam a organização dos dados e a estruturação automática de informações sobre cada ato de pessoal.

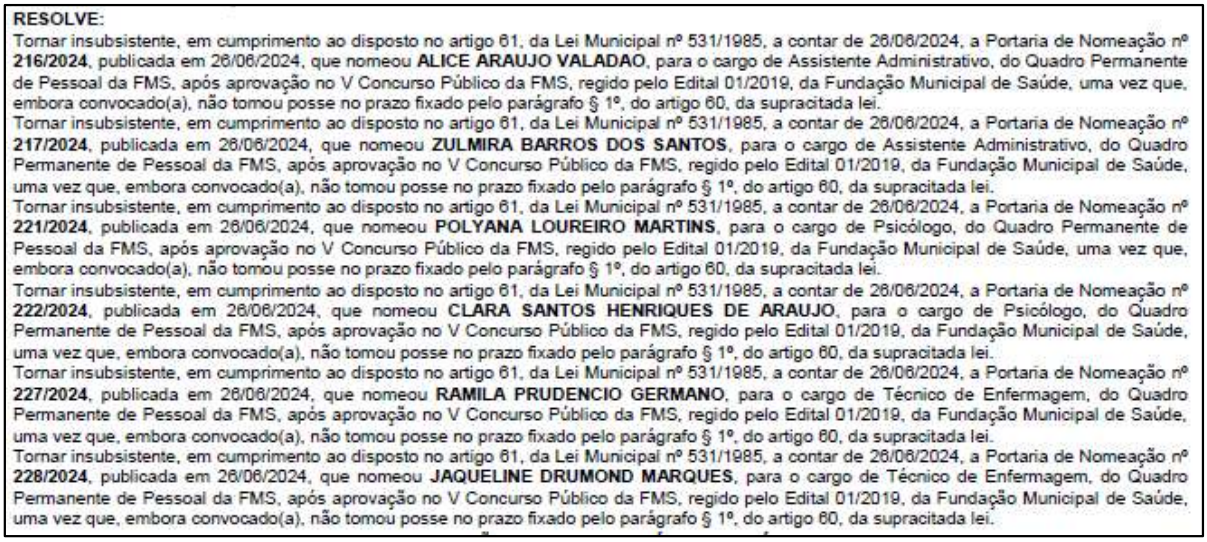
Figura 7 – Exemplo de extração de ato de pessoal em formato JSON.


```
[{"cargo": "Assistente Administrativo", "data": "26-06-
2024", "nome": "ALICE ARAUJO VALADAO", "pagina": 6, "paragrafo": "
Tornar insubsistente em cumprimento ao disposto no artigo 61 da Le
i Municipal nº 5311985 a contar de 26062024 a Portaria de Nomeação
 nº 2162024  publicada em 26062024 que nomeou ALICE ARAUJO VA
LADAO  para o cargo de Assistente Administrativo  do Qua
dro Permanente de Pessoal da FMS após aprovação no V Concu rso Púb
lico da FMS regido pelo Edital 012019 da Fundação Municipal de Saú
de uma vez que embora convocado(a) não tomou posse no prazo fixado
 pelo parágrafo  1º do artigo 60 da supracitada lei.", "tipo"
: "Exoneração"}, {"cargo": "Assistent e Administrativo", "data": "
26-06-
2024", "nome": "ZULMIRA BARROS DOS SANTOS", "pagina": 6, "paragraf
o": "Tornar insubsistente em cumprimento ao disposto no art igo 61
 da Lei Municipal nº 5311985 a contar de 26062024 a Portaria de No
meação nº 2172024  publicada em 26062024 que nomeou ZULMIRA B
ARROS DOS SANTOS  para o cargo de Assistent e Administrativo&
nbsp; do Quadro Permanente de Pessoal da FMS após aprovação no V C
```


Fonte: Elaboração pelos autores.

Na **Figura 8**, há um conteúdo textual em formato PDF, que lista as exonerações de servidores nomeados para diferentes cargos administrativos. O texto segue o padrão de publicações oficiais, mencionando a legislação aplicável e justificando cada exoneração pela ausência de posse no prazo estabelecido. Por exemplo, o nome "ALICE ARAUJO VALADAO", mencionado no JSON com data e cargos específicos, corresponde à primeira exoneração descrita no PDF, confirmando que as informações estruturadas no JSON derivam do conteúdo expresso no documento oficial. Essa relação permite comparar e verificar a consistência dos dados, além de facilitar o processamento automatizado do conteúdo.

Figura 8 – Exemplo de ato de pessoal em Diário Oficial em formato PDF.



RESOLVE:
Tornar insubsistente, em cumprimento ao disposto no artigo 61, da Lei Municipal nº 531/1985, a contar de 26/06/2024, a Portaria de Nomeação nº **216/2024**, publicada em 26/06/2024, que nomeou **ALICE ARAUJO VALADAO**, para o cargo de Assistente Administrativo, do Quadro Permanente de Pessoal da FMS, após aprovação no V Concurso Público da FMS, regido pelo Edital 01/2019, da Fundação Municipal de Saúde, uma vez que, embora convocado(a), não tomou posse no prazo fixado pelo parágrafo § 1º, do artigo 60, da supracitada lei.
Tornar insubsistente, em cumprimento ao disposto no artigo 61, da Lei Municipal nº 531/1985, a contar de 26/06/2024, a Portaria de Nomeação nº **217/2024**, publicada em 26/06/2024, que nomeou **ZULMIRA BARROS DOS SANTOS**, para o cargo de Assistente Administrativo, do Quadro Permanente de Pessoal da FMS, após aprovação no V Concurso Público da FMS, regido pelo Edital 01/2019, da Fundação Municipal de Saúde, uma vez que, embora convocado(a), não tomou posse no prazo fixado pelo parágrafo § 1º, do artigo 60, da supracitada lei.
Tornar insubsistente, em cumprimento ao disposto no artigo 61, da Lei Municipal nº 531/1985, a contar de 26/06/2024, a Portaria de Nomeação nº **221/2024**, publicada em 26/06/2024, que nomeou **POLYANA LOUREIRO MARTINS**, para o cargo de Psicólogo, do Quadro Permanente de Pessoal da FMS, após aprovação no V Concurso Público da FMS, regido pelo Edital 01/2019, da Fundação Municipal de Saúde, uma vez que, embora convocado(a), não tomou posse no prazo fixado pelo parágrafo § 1º, do artigo 60, da supracitada lei.
Tornar insubsistente, em cumprimento ao disposto no artigo 61, da Lei Municipal nº 531/1985, a contar de 26/06/2024, a Portaria de Nomeação nº **222/2024**, publicada em 26/06/2024, que nomeou **CLARA SANTOS HENRIQUES DE ARAUJO**, para o cargo de Psicólogo, do Quadro Permanente de Pessoal da FMS, após aprovação no V Concurso Público da FMS, regido pelo Edital 01/2019, da Fundação Municipal de Saúde, uma vez que, embora convocado(a), não tomou posse no prazo fixado pelo parágrafo § 1º, do artigo 60, da supracitada lei.
Tornar insubsistente, em cumprimento ao disposto no artigo 61, da Lei Municipal nº 531/1985, a contar de 26/06/2024, a Portaria de Nomeação nº **227/2024**, publicada em 26/06/2024, que nomeou **RAMILA PRUDENCIO GERMANO**, para o cargo de Técnico de Enfermagem, do Quadro Permanente de Pessoal da FMS, após aprovação no V Concurso Público da FMS, regido pelo Edital 01/2019, da Fundação Municipal de Saúde, uma vez que, embora convocado(a), não tomou posse no prazo fixado pelo parágrafo § 1º, do artigo 60, da supracitada lei.
Tornar insubsistente, em cumprimento ao disposto no artigo 61, da Lei Municipal nº 531/1985, a contar de 26/06/2024, a Portaria de Nomeação nº **228/2024**, publicada em 26/06/2024, que nomeou **JAQUELINE DRUMOND MARQUES**, para o cargo de Técnico de Enfermagem, do Quadro Permanente de Pessoal da FMS, após aprovação no V Concurso Público da FMS, regido pelo Edital 01/2019, da Fundação Municipal de Saúde, uma vez que, embora convocado(a), não tomou posse no prazo fixado pelo parágrafo § 1º, do artigo 60, da supracitada lei.

Fonte: Elaboração pelos autores.

Vale destacar o nível de compreensão que os LLMs podem alcançar, no qual o texto analisado menciona “nomeação” e mesmo assim, o parágrafo do ato foi classificado corretamente como “exoneração”, sendo que essa ocorrência foi seguida de outras de mesma semântica, as quais foram também classificadas corretamente.

Diferente dos casos apresentados nas **Figura 7** e **Figura 8**, onde o *prompt* solicitado foi executado com sucesso e o ato de pessoal foi classificado corretamente, a **Figura 9** apresenta um caso de erro quando o texto em análise possui restrição de conteúdo, resulta-se numa solicitação recusada.

Figura 9 – Mensagem de erro por restrição.

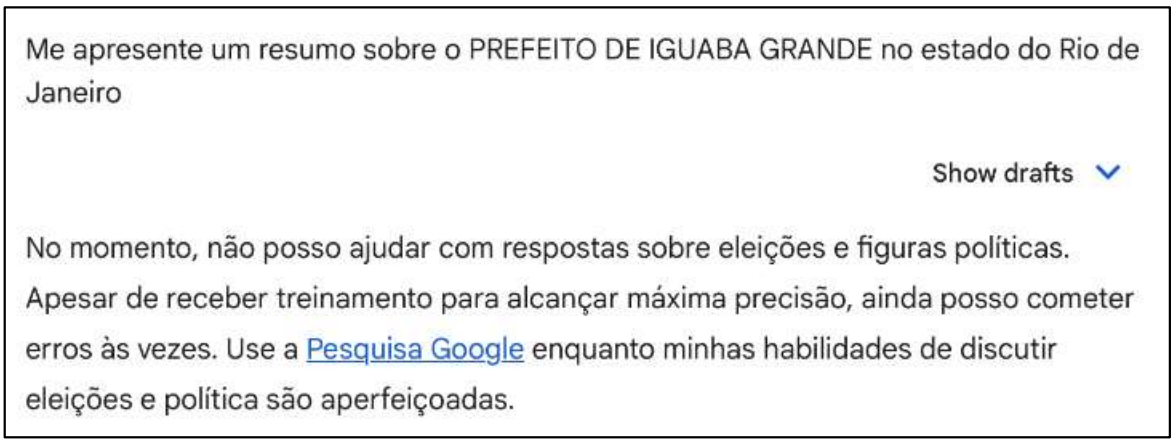
```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-52-5b6f329aba8b> in <cell line: 1>()
     11     else:
     12         print(response_nomeacao.usage_metadata)
---> 13         json_nomeacao = response_nomeacao.text #clean_rsp_json(response_nomeacao.text)
     14         print("Tokens processados (Nomeação): ")
     15         print(response_nomeacao.usage_metadata)

/usr/local/lib/python3.10/dist-packages/google/generativeai/types/generation_types.py in text(self)
    493                 raise ValueError(msg + " Meaning the response was using an unsupported language.")
    494             elif fr is FinishReason.OTHER:
--> 495                 raise ValueError(msg)
    496             elif fr is FinishReason.BLOCKLIST:
    497                 raise ValueError(msg)

ValueError: Invalid operation: The `response.text` quick accessor requires the response to contain a valid `Part`, but none were returned.
```

Fonte: Elaboração pelos autores.

Conforme pôde ser visto na **Figura 9**, o erro descreveu um conteúdo inválido em nível de “response.text” pois identificou-se valor de erro (“ValueError”). Casos como esse podem ocorrer nos modelos Gemini por uma série de motivos, conforme especificados na documentação oficial em estados de “finish_reason” e definidos em execução por critérios de segurança chamados de “safety_settings”.

Para fins de exemplo prático, a **Figura 10** apresenta a aplicação do Gemini, disponível em https://gemini.google.com/app/, com o *prompt* executado contendo o termo “PREFEITO DE IGUABA GRANDE” e o mesmo prompt foi recusado.

Figura 10 – Exemplo de prompt no Google Gemini com solicitação recusada.



Fonte: Elaboração pelos autores.

É importante destacar que, geralmente, as publicações de atos de pessoal contemplam pessoas na forma de figuras políticas (prefeito, governador etc.). De modo similar ao apresentado no prompt, foram percebidos dois casos com erro semelhante durante a execução da amostra de PDF, a qual continha 105 arquivos PDFs.

Por fim, após o procedimento finalizado, o resultado retornado em JSON é convertido ao formato de *Dataframe*, estruturando as informações desde o parágrafo do ato publicado até a classificação do conteúdo deste parágrafo em nomeação ou exoneração. Os atributos extraídos nesse procedimento estão descritos na **Tabela 8**.

Tabela 8 – Descrição dos atributos identificados por ato.

Atributos	Descrição
Filename	Nome do arquivo processado
Nome	Nome da pessoa nomeada ou exonerada
Matrícula	Matrícula da pessoa nomeada ou exonerada, se disponível

Atributos	Descrição
Data	Data de publicação do ato no documento (formato DD-MM-YYYY)
Cargo	Cargo para o qual a pessoa foi nomeada (ou exonerada)
Ato	Tipo de ato: Nomeação ou Exoneração
Parágrafo	Parágrafo de texto do documento onde o ato foi mencionado
Página	Página do documento onde o ato foi identificado

Fonte: Elaborado pelos autores.

3 CONSIDERAÇÕES FINAIS

O projeto **"Desenvolvimento de uma Metodologia para a Coleta e Identificação de Atos Administrativos de Interesse nos Diários Oficiais dos Jurisdicionados do Tribunal de Contas do Estado do Rio de Janeiro (TCE-RJ)"** permitiu a elaboração e experimentação de uma metodologia robusta, explorando técnicas de mineração de dados, processamento de linguagem natural (NLP) e aprendizado de máquina para identificar e classificar atos administrativos de interesse nos Diários Oficiais. O estudo empregou o modelo CRISP-DM como estrutura, possibilitando uma abordagem estruturada e iterativa, que se adaptou às particularidades dos dados e objetivos da pesquisa.

Foram desenvolvidas duas abordagens distintas de classificação: uma com o classificador Random Forest, voltada para dados segmentados e previamente estruturados, e outra com um *Large Language Model* (LLM), capaz de lidar com textos mais complexos e identificar contextos variados, mesmo quando nomeações e exonerações coexistiam no mesmo parágrafo. A análise dos resultados demonstrou que ambas as abordagens possuem altos níveis de precisão e são adequadas para diferentes cenários de uso. O *Random Forest* mostrou-se eficiente para a classificação de dados estruturados, enquanto o LLM apresentou flexibilidade para atuar em contextos linguísticos mais amplos, mantendo a precisão mesmo em estruturas textuais complexas.

Além da precisão alcançada, um dos principais benefícios deste trabalho foi demonstrar a viabilidade do uso de novas tecnologias e ferramentas, como modelos de linguagem em larga escala, na automatização de tarefas repetitivas e no suporte à

análise de dados no setor público. A disponibilização de uma base de dados desestruturada do Diário Oficial, aliada a ferramentas de busca e indexação, como o Bing, também mostrou potencial para agilizar as consultas e análises pelos auditores do TCE-RJ, facilitando a fiscalização e o controle.

Para expandir o alcance da pesquisa, é recomendada a inclusão de uma análise abrangente de outros atos administrativos, como contratos, licitações, convênios e publicações de natureza financeira. Esse crescimento permitiria uma visão mais ampla e detalhada das ações governamentais, favorecendo a identificação de padrões e a detecção de possíveis irregularidades em áreas além das nomeações e exonerações. A metodologia poderia ser adaptada para classificar esses novos tipos de atos, utilizando algoritmos de processamento de linguagem natural (NLP) e aprendizado de máquina (ML) que se ajustem às especificidades terminológicas e contextuais desses documentos.

REFERÊNCIAS

ARAÚJO, Pedro H. Luz de; CAMPOS, Teófilo E. de; SOUSA, Marcelo M. S. de; Inferring the source of official texts: can SVM beat ULMFiT? *In:* INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE (PROPOR), 2020, Évora, Portugal. **Proceedings** [...] Évora, Portugal: Springer, 2-4 mar. 2020. p. 76-86.

BERRAZEGA, Ines *et al.* A knowledge-based approach for provisions' categorization in Arabic normative texts. *In:* SILHAVY, R. *et al.* **Artificial Intelligence Perspectives in Intelligent Systems**. Cham: Springer, 2016. v. 464, p. 415-425. Disponível em: https://doi.org/10.1007/978-3-319-33625-1_37.

BERRAZEGA, Ines *et al.* A semantic annotation model for Arabic legal texts. *In:* HELLENIC CONFERENCE ON ARTIFICIAL INTELLIGENCE (SETN), 9., 2016, Thessaloniki, Greece. **Proceedings** [...]. New York: ACM, 2016. Session: AI Applications, p. 1-8. Disponível em: https://doi.org/10.1145/2903220.2903244.

BERRAZEGA, Ines *et al.* A linguistic method for Arabic normative provisions' annotation based on contextual exploration. *In:* INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION SYSTEMS (ICICS), 7., 2016, Irbid, Jordan. **Proceedings** [...]. New York: IEEE, 5-7 apr. 2016. p. 347–352.

BRANDÃO, Stainam *et al.* Knowledge representation of Brazilian official gazettes for chronological recovery of laws. *In:* CONFERENCE ON INFORMATION SYSTEMS, 2011, Rio de Janeiro. **Proceedings** [...] Rio de Janeiro: IADIS, 5-8 nov, 2011. p. 540–544.

BREIMAN, Leo; FRIEDMAN, Jerome; OLSHEN, R. A.; STONE, Charles J. **Classification and regression trees**. 1st. ed. Boca Raton: Chapman and Hall/CRC, 1984. Disponível em: https://doi.org/10.1201/9781315139470. Acesso em: 22 dez. 2025.

BREIMAN, L. Random forests. **Machine Learning**, [*s. l.*], v. 45, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 22 dez. 2025.

CONSTANTINO, Kattiana *et al.* Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. *In:* SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS (SBBD), 37., 2022, Búzios. **Anais** [...]. Porto Alegre: SBC, 19-23 set. 2022. p. 304–316. Disponível em: https://sol.sbc.org.br/index.php/sbbd/article/view/21815. Acesso em: 23 dez. 2025.

CONSTANTINO, Kattiana *et al.* Using active learning for segmentation and semantic classification of legal acts extracted from official diaries. **Journal of Information and Data Management**, Porto Alegre, v. 14, n. 1, 2023. Disponível em: https://doi.org/10.5753/jidm.2023.3181. Acesso em: 23 dez. 2025.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: pre-training of deep bidirectional transformers for language understanding. *In:* CONFERENCE OF THE

NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2019, Minneapolis. **Proceedings** [...]. Minneapolis: Association for Computational Linguistics, 2019. p. 4171–4186.

GE, Yingqiang *et al.* OpenAGI: when LLM meets domain experts. **Advances in Neural Information Processing Systems**, v. 36, 2024. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets_and_Benchmarks.pdf. Acesso em: 23 dez. 2025.

GREFENSTETTE, G. Tokenization. *In:* VAN HALTEREN, H. (ed.). **Syntactic wordclass tagging**. Dordrecht: Springer, 1999. p. 117–133. Disponível em: https://doi.org/10.1007/978-94-015-9273-4_9.

GUIMARÃES, Gabriel M. C. *et al.* DODFMiner: an automated tool for named entity recognition from official gazettes. **Neurocomputing**, London, v. 568, p. 1–10, feb. 2024. Disponível em: https://doi.org/10.1016/j.neucom.2023.127064.

JI, Ziwei *et al.* Towards mitigating LLM hallucination via self reflection. *In:* FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EMNLP 2023), 2023, Singapure. **Proceedings** [...]. Kerrville, TX: Association for Computational Linguistics, 6-10 dec. 2023. p. 1827-1843. Disponível em: https://aclanthology.org/2023.findings-emnlp.123.pdf. Acesso em: 23 dez. 2025.

KAVLAKOGLU, Eda. **O que é random forest?.** Tradução de What is random forest?. New York: São Paulo: IBM Research, 25 jul. 2024. Disponível em: https://www.ibm.com/br-pt/topics/random-forest. Acesso em: 25 jul. 2024.

NEVES JUNIOR, R. B. das; MELO, W. F. D. M.; FAGUNDES, R. A. D. A.; MACIEL, A. M. A. Extração de informação e mineração de dados no diário oficial de Pernambuco. **REPE: Revista de Engenharia e Pesquisa Aplicada**, Pernambuco, v. 3, n. 3, p. 107-113, 2018. Disponível em: http://revistas.poli.br/index.php/repa/article/view/892/449. Acesso em: 5 dez. 2025.

PINTO, Fernando A. D. G.; LIFSCHITZ, Sérgio; HAEUSLER, Edward H. A knowledge base of public acts based on the grammar of the official gazette. *In:* INTERNATIONAL CONFERENCE ON DIGITAL GOVERNMENT TECHNOLOGY AND INNOVATION (DGTi-CON), 2022. **Proceedings** [...]. Bangkok, Thailand: IEEE, 24-25 mar. 2022. p. 24–29. Disponível em: https://doi.org/10.1109/DGTi-CON53875.2022.9849196. Acesso em: 22 dez. 2025.

PINTO, Fernando A. D. G.; HAEUSLER, Edward H.; LIFSCHITZ, Sérgio. Transparência pública automatizada a partir da gramática do diário oficial. *In:* WORKSHOP DE COMPUTAÇÃO APLICADA EM GOVERNO ELETRÔNICO (WCGE 2021), 9., 2021. **Anais eletrônicos** [...]. Disponível em: https://sol.sbc.org.br/index.php/wcge/article/view/15977/15818. Acesso em: 5 dez. 2025.

ROCHA, João Paulo L. **Inteligência de fontes abertas**: um estudo de caso sobre descoberta de conhecimento no diário oficial da união. 2011. Dissertação (Mestrado

em Informática) – Universidade Católica de Brasília, Brasília. Disponível em: https://bdtd.ucb.br:8443/jspui/handle/123456789/1336. Acesso em: 5 dez. 2025.

RODRÍGUEZ, Marcia M.; BEZERRA, Byron L. D. Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias). **REPE**: **Revista de Engenharia e Pesquisa Aplicada**, Pernambuco, v. 5, n. 1, p. 67-77, 2020. Disponível em: http://revistas.poli.br/index.php/repa/article/view/1204. Acesso em: 5 dez. 2025.

OPEN KNOWLEDGE BRASIL. **Querido Diário**. [*S. l.*]: OKBR, 2024. Disponível em: https://queridodiario.ok.org.br/sobre. Acesso em: 5 dez. 2025.

PEDREGOSA, Fabian *et al.* Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, [*s. l.*], v. 12, n. 8, p. 2825–2830, 2011. Disponível em: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: 23 dez. 2025.

OGAWA, Yasuhiro *et al.* Extraction of legal bilingual phrases from the Japanese official gazette, English edition. *In:* INTERNATIONAL CONFERENCE ON KNOWLEDGE AND SYSTEMS ENGINEERING (KSE), 8., 2016, Hanoi. **Proceedings** [...]. New York: IEEE, 6-8 oct. 2016. p. 258–263.

SANH, Victor; DEBUT, Lysandre; CHAUMOND, Julien; WOLF, Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *In:* CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 33., 2019, Vancouver, Canada. **Proceedings** [...]. Vancouver, Canada: EMC2, 9-13 dec. 2019. Disponível em: https://www.emc2-ai.org/assets/docs/neurips-19/emc2-neurips19-paper-33.pdf. Acesso em: 23 dez. 2025.

SHEARER, C. The CRISP-DM model: the new blueprint for data mining. **Journal of Data Warehousing**, [*s. l.*], v. 5, n. 4, p. 13-22, 2000.

XAVIER, Bruno D.; SILVA, Alcione Dias da; GOMES, Georgia R. G. Uma arquitetura híbrida para a indexação de documentos do diário oficial do município de Cachoeiro de Itapemirim. **Transinformação**, Campinas, v. 27, n. 1, p. 83-95, jan./abr. 2015. Disponível em: https://periodicos.puc-campinas.edu.br/transinfo/article/view/6056. Acesso em: 5 dez. 2025.

VASWANI, Aahish *et al.* Attention is all you need. *In:* CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2017), 31., 2017, Long Beach, CA. **Proceedings** [...]. San Diego, CA: NeurIPS, 4-9 dec. 2017. Disponível em: https://arxiv.org/pdf/1706.03762. Acesso em: 23 dez. 2025.

Sobre os autores

Wellington Souza Amaral | *e-mail*: wellingtonsa@tcerj.tc.br
Mestre em Ciência da Computação – Centro Federal de Educação Tecnológica do Rio de Janeiro. Auditor de Controle Externo no Tribunal de Contas do Estado do Rio de Janeiro.

https://orcid.org/0009-0000-0389-6656
http://lattes.cnpq.br/0015927168023271

Gustavo Alexandre Sousa Santos | *e-mail*: gassantos@id.uff.br
Mestre em Ciência da – Centro Federal de Educação Tecnológica do Rio de Janeiro. Assessor no Tribunal de Contas do Estado do Rio de Janeiro.

https://orcid.org/0000-0002-3604-9194
http://lattes.cnpq.br/6269223842813109

Eduardo Bezerra da Silva | *e-mail*: ebezerra@cefet-rj.br
Doutor em Engenharia de Sistemas e Computação – Universidade Federal do Rio de Janeiro. Professor titular da Escola de Informática e Computação do Centro Federal de Educação Tecnológica do Rio de Janeiro.

https://orcid.org/0000-0001-9177-5503
http://lattes.cnpq.br/7568520840965379

Leonardo Silva de Lima | *e-mail*: leonardo.delima@ufpr.br
Doutor em Engenharia de Produção – Universidade Federal do Rio de Janeiro. Professor da Universidade Federal do Paraná.

https://orcid.org/0000-0002-4949-7850
http://lattes.cnpq.br/0206233750299857

Augusto César Benvenuto de Almeida | *e-mail*: augustocba@tcerj.tc.br
Graduado em Engenharia da Computação – Universidade Federal de Pernambuco. Auditor de Controle Externo no Tribunal de Contas do Estado do Rio de Janeiro.

https://orcid.org/0009-0005-2768-8620
http://lattes.cnpq.br/4732180899245369