PGCOMP - Programa de Pós-Graduação em Ciência da Computação
Universidade Federal da Bahia (UFBA)
Av. Milton Santos, s/n - Ondina
Salvador, BA, Brasil, 40170-110

https://pgcomp.ufba.br
pgcomp@ufba.br

No contexto das compras governamentais no Brasil, a eficiência e o monitoramento contínuo dos gastos representam desafios significativos para a gestão pública. O Governo Federal do Brasil, em 2023, emitiu um total de 1.761.910 de notas fiscais para diversos tipos de aquisições, resultando em um montante de R$ 76,62 bilhões em negociações com entidades privadas (TRANSPARÊNCIA, 2024). Essas aquisições governamentais abrangem um amplo espectro da máquina pública, incluindo aquisições para compras de materiais utilizados em construção de rodovias, manutenção de escolas e hospitais, utilização de bens pelo funcionalismo público, dentre outros fins. As aquisições desses insumos são distribuídas em diversos locais do território nacional, gerando um volume crescente e diversificado de informações, presentes em contratos e notas fiscais de produtos e serviços.

No entanto, essas compras governamentais são frequentemente um campo fértil para a ocorrência de conluios e fraudes (OECD, 2007), como superfaturamento nos preços dos produtos, monopólios dos fornecedores, propina para agentes públicos, etc.

O objetivo deste trabalho é comparar o desempenho de modelos de Processamento de Linguagem Natural (PLN) na tarefa de detecção - com base no extrato das compras governamentais - de empresas que já foram punidas por órgãos governamentais, como a Controladoria-Geral da União (CGU). Os dados utilizados são públicos e periodicamente atualizados através do portal de Dados Abertos do Governo Federal.

Os resultados deste trabalho mostram que é possível utilizar modelos de linguagem natural como uma pré-etapa de investigação de compras suspeitas, fornecendo uma classificação de compras potencialmente problemáticas e que posteriormente podem ser avaliadas por um especialista, dessa forma, reduzindo a carga de trabalho humana ao reduzir a lista de compras para uma quantidade menor e mais focalizada.

Palavras-chave: Grandes Modelos de Linguagem, Análise de contratos públicos, Transparência pública, Detecção de fraude, Transformadores.

MSC | 193 | 2025

Detecção de empresas potencialmente não confiáveis por meio de extratos de compras governamentais: uma aplicação com modelos de linguagem natural

Cleiton Otavio da Exaltação Rocha

UFBA

Detecção de empresas potencialmente não confiáveis por meio de extratos de compras governamentais: uma aplicação com modelos de linguagem natural

Cleiton Otavio da Exaltação Rocha

Dissertação de Mestrado

Universidade Federal da Bahia

Programa de Pós-Graduação em
Ciência da Computação

Junho | 2025

Universidade Federal da Bahia
Instituto de Computação

Programa de Pós-Graduação em Ciência da Computação

DETECÇÃO DE EMPRESAS POTENCIALMENTE NÃO CONFIÁVEIS POR MEIO DE EXTRATOS DE COMPRAS GOVERNAMENTAIS: UMA APLICAÇÃO COM MODELOS DE LINGUAGEM NATURAL

Cleiton Otavio da Exaltação Rocha

DISSERTAÇÃO DE MESTRADO

Salvador
05 de junho de 2025

CLEITON OTAVIO DA EXALTAÇÃO ROCHA

DETECÇÃO DE EMPRESAS POTENCIALMENTE NÃO CONFIÁVEIS POR MEIO DE EXTRATOS DE COMPRAS GOVERNAMENTAIS: UMA APLICAÇÃO COM MODELOS DE LINGUAGEM NATURAL

Este Projeto de Dissertação foi apresentado ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Gecynalda Soares da Silva Gomes

Salvador
05 de junho de 2025

Sistema de Bibliotecas - UFBA

R672 Rocha, Cleiton Otavio da Exaltação.
Detecção de empresas potencialmente não confiáveis por meio de extratos de compras governamentais: uma aplicação com modelos de linguagem natural / Cleiton Otavio da Exaltação Rocha – Salvador, 2025.
61p.: il.

Orientadora: Profa. Dra. Gecynalda Soares da Silva Gomes.
Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Computação, 2025.

1. Modelos de Linguagem. 2. Contratos públicos - Análise. 3. Transparência pública. I. Gomes, Gecynalda Soares da Silva. II. Universidade Federal da Bahia. Instituto de Computação. III Título.

CDU – 004.4

Termo de Aprovação

Cleiton Otavio da Exaltação Rocha

Detection of potentially untrustworthy companies through government procurement extracts: an application with natural language models.

Esta Dissertação foi julgada adequada à obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da UFBA.

Salvador, 05 de junho de 2025

gov.br Documento assinado digitalmente
GECYNALDA SOARES DA SILVA GOMES
Data: 06/06/2025 12:31:52-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Gecynalda Soares da Silva Gomes
(Orientadora - PGCOMP)

gov.br Documento assinado digitalmente
MARLO VIEIRA DOS SANTOS E SOUZA
Data: 10/06/2025 19:59:23-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marlo Vieira dos Santos e Souza
(IC/UFBA)

gov.br Documento assinado digitalmente
RICARDO FERREIRA DA ROCHA
Data: 06/06/2025 13:32:31-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ricardo Ferreira da Rocha

(IME/UFBA)

RESUMO

No contexto das compras governamentais no Brasil, a eficiência e o monitoramento contínuo dos gastos representam desafios significativos para a gestão pública. O Governo Federal do Brasil, em 2023, emitiu um total de 1.761.910 de notas fiscais para diversos tipos de aquisições, resultando em um montante de R$ 76,62 bilhões em negociações com entidades privadas (TRANSPARÊNCIA, 2024). Essas aquisições governamentais abrangem um amplo espectro da máquina pública, incluindo aquisições para compras de materiais utilizados em construção de rodovias, manutenção de escolas e hospitais, utilização de bens pelo funcionalismo público, dentre outros fins. As aquisições desses insumos são distribuídas em diversos locais do território nacional, gerando um volume crescente e diversificado de informações, presentes em contratos e notas fiscais de produtos e serviços.

No entanto, essas compras governamentais são frequentemente um campo fértil para a ocorrência de conluios e fraudes (OECD, 2007), como superfaturamento nos preços dos produtos, monopólios dos fornecedores, propina para agentes públicos, etc.

O objetivo deste trabalho é comparar o desempenho de modelos de PLN na tarefa de detecção - com base no extrato das compras governamentais - de empresas que já foram punidas por órgãos governamentais, como a CGU. Os dados utilizados são públicos e periodicamente atualizados através do portal de Dados Abertos do Governo Federal.

Os resultados deste trabalho mostram que é possível utilizar modelos de linguagem natural como uma pré-etapa de investigação de compras suspeitas, fornecendo uma classificação de compras potencialmente problemáticas e que posteriormente podem ser avaliadas por um especialista, dessa forma, reduzindo a carga de trabalho humana ao reduzir a lista de compras para uma quantidade menor e mais focalizada.

Palavras-chave: Grandes Modelos de Linguagem, Análise de contratos públicos, Transparência pública, Detecção de fraude, Transformadores.

ABSTRACT

In the context of government procurement in Brazil, efficiency and continuous monitoring of expenditure represent significant challenges for public management. In 2023, the Brazilian Federal Government issued a total of 1,761,910 invoices for various types of acquisitions, resulting in an amount of R$76.62 billion in negotiations with private entities (TRANSPARÊNCIA, 2024). These government procurements cover a broad spectrum of the public sector, including procurement for the purchase of materials used in the construction of highways, maintenance of schools and hospitals, use of goods by public servants, among other purposes. The acquisition of these inputs is distributed in various locations throughout the country, generating a growing and diverse volume of information, present in contracts and invoices for products and services.

However, these government procurements are often a fertile ground for the occurrence of collusion and fraud (OECD, 2007), such as overpricing of products, supplier monopolies, bribes to public officials, etc.

The objective of this work is to compare the performance of *Natural Language Processing* (NLP) models in the task of detecting - based on extracts of government purchases - companies that have already been punished by government agencies, such as CGU. The data used are public and periodically updated through the Federal Government's Open Data portal.

The results of this work show that it is possible to use natural language models as a pre-stage of investigation of suspicious purchases, providing a classification of potentially problematic purchases that can later be evaluated by an expert, thus reducing the human workload by reducing the purchase list to a smaller and more focused amount.

Keywords: Large Language Models, Public procurement analysis, Public transparency, Fraud detection, Transformers.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

Capítulo 1

INTRODUÇÃO

As compras governamentais envolvem a aquisição de bens e serviços pelo setor público, com o objetivo de manter as operações da administração pública e proporcionar serviços à população em áreas como educação, saúde, segurança, energia e infraestrutura (KASHAP, 2004). Entre janeiro de 2022 e junho de 2024, o Governo Federal do Brasil emitiu um total de 4,3 milhões de notas fiscais, culminando em R$ 198,33 bilhões gastos na compra de diversos produtos (TRANSPARÊNCIA, 2024).

De acordo com Ribeiro e Júnior (2019), o processo de compras governamentais pode ser segmentado em três fases essenciais: i) a determinação do momento e das especificidades dos bens ou serviços que necessitam ser adquiridos (planejamento de compras); ii) a elaboração de um contrato de aquisição, que envolve, primordialmente, a seleção do parceiro contratual e a definição dos termos sob os quais os produtos ou serviços devem ser fornecidos; e iii) a administração do contrato, visando garantir a sua execução eficaz. Ademais, Ribeiro et al. (2018) ressaltam uma política governamental que estabelece objetivos, tais como a eficiência (efetuar compras pelo menor preço, de maneira oportuna, evitando fraudes e desperdício dos recursos públicos), a promoção da indústria local e a geração de empregos.

No contexto das aquisições governamentais no Brasil, tanto a eficiência quanto o monitoramento contínuo dos gastos representam desafios significativos para a administração pública (COSTA; HOLLNAGEL; BUENO, 2019). Segundo Basdevant et al. (2022), a incidência de fraudes e superfaturamento nas compras públicas, elemento crucial para a execução dos orçamentos governamentais, pode acarretar consequências financeiras relevantes, resultando em déficits mais elevados e na redução do crescimento econômico devido à qualidade inadequada e/ou quantidade insuficiente de infraestrutura. Nesse sentido, Costa, Hollnagel e Bueno (2019) aponta que o sistema vigente prioriza o cumprimento de formalidades legais em detrimento da obtenção de resultados práticos e econômicos, o que gera processos burocráticos, prolongados e frequentemente desarticulados. De acordo com esses autores, a carência de planejamento eficaz nas aquisições públicas intensifica o desperdício de recursos, seja por superestimação de custos ou pela inadequação nas contratações.

Ademais, Plaček et al. (2020) destaca que as entidades governamentais desempenham um papel crucial na prevenção de compras públicas potencialmente inadequadas, exercendo essa função por meio da regulamentação e supervisão de empresas fornecedoras de serviços, bem como dos próprios órgãos governamentais. Considerando o elevado volume de aquisições governamentais realizadas mensalmente e a dificuldade inerente à detecção de irregularidades, torna-se imperativo o desenvolvimento de sistemas de triagem inicial que consigam identificar, de forma automática e eficiente, empresas altamente suspeitas numa fase inicial do vínculo com a administração pública. Tal medida visa não apenas evitar possíveis irregularidades futuras, mas também facilitar o processo de monitoramento e acompanhamento, de modo a identificar compras e aquisições problemáticas.

1.1 PROBLEMA

As compras governamentais são parte importante do processo de manutenção da máquina pública. Segundo Braz et al. (2024), para as empresas privadas, participar de licitações públicas representa uma excelente oportunidade de crescimento, já que o setor público pode adquirir bens garantindo uma receita estável. Segundo Mohallem e Ragazzo (2017), com dados da Transparência Internacional (2014), os gastos com compras públicas somam de 13 a 20% do produto interno bruto mundial, isso evidencia o tamanho e a importância desse mercado.

Dada sua amplitude, este setor também enfrenta desafios; conforme Costa et al. (2022), o elevado volume de dados gerado por essas transações, combinado com a linguagem frequentemente técnica e complexa dos produtos envolvidos, dificulta a análise manual e detalhada pelos auditores e especialistas responsáveis pela transparência e integridade administrativa. Além disso, a presença de empresas com histórico de sanções participando de compras públicas reforça a necessidade de ferramentas automatizadas para a detecção destas situações.

Conforme demonstrado na Tabela 1.1, o número de descrições de itens em compras governamentais supera a marca de 1 milhão entre os anos de 2022 e 2024. Estas descrições abrangem um pouco mais de 7.000 NCMs (Nomenclatura Comum do Mercosul), um código de oito dígitos que identifica mercadorias em países que fazem parte do Mercado Comum do Sul (Mercosul). Em paralelo, a Tabela 1.2 evidencia que, embora a maioria das empresas participantes não possua registros de sanções, uma parcela significativa ainda inclui empresas com algum tipo de sanção (1.440 empresas em 2022 e 1.115 em 2024). Esse contexto destaca a necessidade de ferramentas robustas capazes de lidar com grandes volumes de dados e fornecer suporte na identificação de empresas potencialmente não confiáveis.

Tabela 1.1 Total de NCMs e descrições, por ano

Ano	Total de NCMs	Total de descrições
2022	7.835	1.300.707
2023	7.597	1.303.256
2024	7.400	1.091.357

Fonte: Elaboração própria.

Tabela 1.2 Número de empresas que participaram de compras governamentais, por ano

Classe	2022	2023	2024
Sem registros de sanção	75.969	85.745	92.820
Com alguma sanção (passada ou presente)	1.440	1.225	1.115

Fonte: Elaboração própria.

Embora avanços tecnológicos recentes tenham possibilitado o uso de ferramentas baseadas em Inteligência Artificial (IA), especialmente em modelos de PLN, para a análise automatizada de grandes volumes de dados textuais, a aplicação efetiva dessas tecnologias em documentos relacionados às compras públicas ainda é limitada. Muitas soluções existentes falham em lidar com nuances linguísticas, como expressões regionais, ambiguidades contextuais ou vocabulário técnico, resultando em alta taxa de falsos positivos ou falsos negativos na detecção de irregularidades.

1.2 OBJETIVO

O presente estudo visa avaliar a viabilidade da aplicação de modelos de PLN na tarefa de detecção, baseada no extrato das compras governamentais, de empresas que já tenham sido sancionadas por órgãos governamentais, como a Controladoria-Geral da União (CGU). A classificação é conduzida de forma binária: 0 e 1, onde "0" corresponde às compras efetuadas com empresas nunca punidas no âmbito público e "1" refere-se às compras efetuadas com empresas que já foram sujeitas a algum tipo de sanção. No processo de classificação, foram empregadas três técnicas: *zero-shot*, *few-shots* e *fine-tuning*. Após a classificação, diversas métricas de desempenho foram utilizadas para prover uma comparação entre os modelos, culminando na seleção de um modelo final, o qual foi submetido a uma análise mais detalhada de seus resultados.

Além da análise da viabilidade do uso de modelos de PLN para a identificação de empresas previamente penalizadas por meio de aquisições governamentais, também constituem objetivos deste estudo:

1. Analisar os dados de compras governamentais para entender quais produtos e quais empresas mais negociaram com o Governo Federal.

2. Implementar e disponibilizar o modelo com melhor desempenho nos testes para realizar a detecção de empresas em novos dados disponíveis periodicamente.

3. Comparar o desempenho dos modelos implementados neste trabalho com outros modelos utilizados na literatura, estabelecendo critérios objetivos de medição.

1.3 ORGANIZAÇÃO

O presente texto está estruturado em cinco capítulos, com exceção da introdução atual: Fundamentação Teórica, Trabalhos Relacionados, Detecção de Empresas Inidôneas em Compras Governamentais, Resultados e Considerações Finais. No capítulo de Fundamentação Teórica, realiza-se uma discussão aprofundada sobre o setor de compras gover-

namentais, abordando seus objetivos, abrangência e desafios, com ênfase na identificação de irregularidades em aquisições, além de introduzir uma explicação sobre como funciona a arquitetura base para os modelos utilizados neste trabalho. No capítulo de Trabalhos Relacionados, efetua-se uma revisão de estudos conduzidos por pesquisadores brasileiros e internacionais no campo da inteligência computacional aplicada a dados públicos, investigando irregularidades como superfaturamento, cartéis empresariais e aquisições irregulares, entre outros aspectos. O capítulo sobre Detecção de Empresas Inidôneas em Compras Governamentais descreve detalhadamente todos os procedimentos adotados para a coleta e estruturação dos dados, os recursos computacionais utilizados para o desenvolvimento do trabalho, os modelos implementados e testados, bem como a disponibilização dos códigos. No capítulo de Resultados tem-se a análise exploratória dos dados, o desempenho do modelo tanto durante o treinamento quanto com novos dados, demonstrando sua capacidade de generalização e eficácia ao atingir métricas de desempenho que o qualificam para a detecção de aquisições possivelmente irregulares. Finalmente, as Considerações Finais expõem as principais conclusões, discutem as potenciais limitações deste estudo e sugerem direções para pesquisas futuras.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2.1 COMPRAS GOVERNAMENTAIS

2.1.1 Introdução e conceitos básicos

O processo de compras governamentais corresponde à contratação de bens e serviços por instituições públicas, com o propósito de suprir as demandas da administração e garantir a oferta de serviços à coletividade. Este escopo inclui desde bens de consumo imediato até serviços complexos e infraestrutura. Consoante a Organização das Nações Unidas (ONU), tal processo abarca todas as fases, desde a identificação das necessidades até a formalização de contratos (RIBEIRO; JúNIOR, 2019). Ademais, conforme Ribeiro e Júnior (2019), o mercado brasileiro de compras governamentais corresponde a aproximadamente 12,5% do PIB nacional, uma média calculada para o período de 2006 a 2016. Este percentual evidencia o considerável poder econômico do Estado como consumidor, que movimenta cerca de R$ 500 bilhões anualmente, englobando todos os entes federativos.

O mercado de compras públicas no Brasil segue um arcabouço normativo complexo, centrado principalmente na Lei nº 8.666/1993 e em regulamentações complementares, como a Lei do Pregão Eletrônico (Lei nº 10.520/2002). Essas normas visam garantir a legalidade, a transparência e a eficiência dos processos (COSTA; HOLLNAGEL; BUENO, 2019).

A Lei nº 8.666/1993, também conhecida como Lei de Licitações e Contratos, estabeleceu as bases para os processos de aquisição pública no Brasil, definindo normas gerais aplicáveis a todas as esferas de governo. Com o objetivo de assegurar a legalidade, a transparência e a eficiência das contratações públicas, a Lei nº 8.666/1993 adota princípios como a isonomia, a publicidade, a economicidade e a moralidade. Essa estrutura normativa busca garantir que a escolha das propostas mais vantajosas para a administração pública ocorra de maneira imparcial e ética. Para isso, a lei detalha critérios de julgamento, que variam entre o menor preço, a melhor técnica ou uma combinação de ambos, dependendo da natureza do contrato. Por fim, contempla dispositivos relativos a sanções e penalidades, destinadas a garantir a conformidade dos contratos administrativos e a punir práticas inadequadas, como descumprimento contratual ou fraudes. Ela também

trata de diferentes modalidades, como concorrência, tomada de preços, convite, concurso e leilão, cada uma delas voltada para situações específicas de contratação, dependendo do valor, da natureza do objeto e das peculiaridades do processo.

A Lei nº 10.520/2002, que introduziu o pregão como uma modalidade de licitação, foi desenvolvida com o objetivo de modernizar e simplificar os processos de aquisição de bens e serviços comuns. Em contraste com as modalidades estabelecidas pela Lei nº 8.666/1993, o pregão proporciona maior rapidez e competitividade, possibilitando lances sucessivos pelos licitantes, o que intensifica a busca pela proposta mais vantajosa. Esta modalidade, particularmente em sua forma eletrônica, destaca-se pela inovação no uso de plataformas digitais, que ampliam o alcance do processo licitatório, reduzem custos e elevam a transparência. Tal estrutura ágil é caracterizada por prazos reduzidos para a apresentação de propostas e recursos, além de um formato em que as ofertas são julgadas inicialmente, com a verificação da documentação do licitante vencedor ocorrendo posteriormente.

Esses dois marcos legais são complementares e refletem diferentes necessidades no âmbito das compras públicas. A Lei nº 8.666/1993 permanece como referência para contratações mais complexas, enquanto a Lei nº 10.520/2002 atende à crescente demanda por agilidade e eficiência nas aquisições rotineiras. Juntas, elas configuram um sistema normativo robusto, mas também desafiador, dado o volume de regulamentações e a constante necessidade de aprimoramento para lidar com a evolução tecnológica e as demandas de transparência e eficiência na administração pública.

Complementarmente, segundo Costa, Hollnagel e Bueno (2019), ainda que o pregão eletrônico seja a modalidade predominante, especialmente no âmbito federal, a concorrência ainda é amplamente usada em contextos municipais e em grandes obras devido à sua adaptabilidade a diferentes níveis de complexidade.

2.1.2 Modalidades de licitação vigentes no Brasil

As formas de licitação atualmente em vigor no Brasil foram instituídas pela Lei nº 8.666/93, conforme estabelecido:

- Concorrência - nesta modalidade, todos os interessados que comprovem preencher os requisitos descritos no edital podem participar. O anúncio de notificação pública será amplamente divulgado e os bens serão adquiridos por meio de licitação competitiva, com valores monetários altamente estimados;

- Convite - esta modalidade é realizada entre inscritos ou não, selecionados e convidados, sendo o número mínimo de convocados definido pela unidade administrativa igual a três. A data, local e documento de solicitação são definidos sem anúncio público. É importante entender que este modo é projetado para itens de baixo preço;

- Tomada de preços - as aquisições de mercadorias de valor intermediário são feitas por tomada de preços, que pode ser substituída por outras modalidades, como concurso ou convite, desde que os interessados estejam cadastrados e cumpram

as exigências de notificação até o terceiro dia anterior ao dia do recebimento das ofertas;

- Concurso - as partes envolvidas podem ser qualquer pessoa com interesse nas artes, ciências ou empreendimentos técnicos e desejando conceder reconhecimento monetário ou outras formas de reconhecimento aos mais merecedores;

- Leilão - nesta modalidade, qualquer pessoa com interesse investido pode vender itens ou coisas que foram confiscadas ou penhoradas pelo governo, mas não são úteis para o Estado.

De acordo com Meirelles (2010), a dispensa de licitação é justificável exclusivamente em situações de emergência devidamente reconhecidas e declaradas, com a finalidade de corrigir uma anomalia ou evitar danos. Durante ocorrências de guerra, grave perturbação da ordem ou calamidade pública, pode-se conceder autorização para a dispensa de licitação na área impactada.

Por fim, o pregão eletrônico constitui uma sexta modalidade de licitação, regulamentada pela Lei nº 10.520/02, que, conforme Freitas e Maldonado (2013):

> "O pregão eletrônico caracteriza-se pela utilização de recursos de tecnologia da informação nos procedimentos licitatórios, proporcionando a comunicação e a interação à distância, pela internet, entre os agentes públicos responsáveis pela licitação (pregoeiro e equipe de apoio) e os licitantes (empresas interessadas em fornecer ou contratar com a Administração). Destacam-se entre as vantagens proporcionadas por este instrumento a maior celeridade dos procedimentos, a ampliação do leque de interessados e a maior transparência e publicidade dos atos administrativos" (FREITAS; MALDONADO, 2013).

Nesta modalidade, as propostas e lances são feitos em uma sessão aberta, presencial ou online, sendo selecionada a proposta ou oferta de maior e melhor valor para o contratante.

2.1.3 Empresas inidôneas em processos de compras públicas

Conforme estipulado pela Constituição, é imperativo que todas as aquisições e contratos governamentais sejam submetidos a um processo licitatório. Tal procedimento consiste em permitir que empresas concorram pela oferta de serviços ou pela venda de produtos que serão utilizados pela Administração Pública. De acordo com Ishikawa e Alencar (2020), a insuficiência de protocolos adequados torna os processos licitatórios vulneráveis a fraudes.

Ainda segundo Ishikawa e Alencar (2020), isso ocorre em virtude dos seguintes fatores:

- Superfaturamento de Preços, que ocorre quando valor do item licitado é maior que o praticado no mercado. Esse problema pode ocorrer devido a pesquisa de preços deficiente durante a elaboração do termo de referência.

- Falhas no processo competitivo, como conluio entre empresas para evitar uma concorrência real de preços.

- Superestimativa de Quantidades: As licitações podem especificar quantidades maiores do que o necessário. Como, por exemplo, a inclusão de mais materiais do que o exigido em obras similares.

- Exigências Excessivas de Critérios Qualitativos: Insere-se a exigência de características ou qualidades desnecessárias nos materiais ou serviços. Isso reduz a competição ao limitar os concorrentes aptos, favorecendo fornecedores específicos.

- Brechas nos Editais: Os prazos curtos entre a publicação dos editais e o julgamento das propostas dificultam a fiscalização preventiva. Isso permite a existência de falhas ou "brechas" nos editais que podem ser usadas para viabilizar fraudes.

.

Adicionalmente, estes autores destacam que tais problemas são intensificados por um monitoramento limitado, com escasso uso de ferramentas preventivas, e pela dificuldade em gerir um grande volume de transações públicas.

Segundo Costa et al. (2022), as empresas utilizam diversos métodos para fraudar os processos licitatórios, explorando vulnerabilidades estruturais e operacionais no sistema de compras públicas. As principais estratégias são:

- Uso de sócios em comum: Empresas diferentes participam de uma mesma licitação, mas compartilham sócios. Isso cria a aparência de competição, mas na realidade as propostas são coordenadas para direcionar o resultado.

- Informações de contato em comum: Empresas que participam de licitações apresentam os mesmos números de telefone, endereços ou e-mails, indicando vínculos entre os licitantes que deveriam ser independentes.

- Participação de empresas com irregularidades cadastrais: Licitantes que estão com CNPJ inativo, sancionados (listados no Cadastro de Empresas Inidôneas e Suspensas - CEIS) ou mesmo inexistentes participam e, às vezes, vencem processos licitatórios.

- Empresas vencedoras e perdedoras frequentes: Padrões anômalos indicam que algumas empresas vencem licitações de forma sistemática, enquanto outras frequentemente participam apenas para simular concorrência e perder propositalmente. Tais padrões podem indicar a existência de acordos prévios entre os participantes.

- Empresas que licitam antes do início de suas atividades: Em alguns casos, há registro de empresas participando de processos licitatórios antes mesmo de sua formalização ou registro oficial.

- Direcionamento do edital: Algumas licitações são desenhadas de forma a favorecer determinadas empresas, utilizando exigências específicas que limitam a concorrência. Essas práticas podem incluir cláusulas excessivamente restritivas ou condições que apenas certas empresas conseguem atender.

- Vínculos entre licitantes e agentes públicos: Fraudes podem envolver conexões entre os proprietários ou sócios das empresas participantes e os agentes públicos responsáveis pelo processo, criando um cenário de conflito de interesses.

Como forma de coibir o avanço de fraudes em compras governamentais, diversas ferramentas já estão sendo empregadas em todo o mundo, tanto por governos quanto por empresas, visando otimizar as atividades públicas e privadas (ISHIKAWA; ALENCAR, 2020). Um exemplo disso é o governo dos Estados Unidos, que utilizou técnicas avançadas de análise de dados (analytics) para combater fraudes fiscais, incluindo a detecção de usuários com maior propensão a práticas ilícitas, a identificação de possíveis fraudes anteriores e posteriores ao pagamento, o monitoramento de ameaças internas e o uso de relatórios analíticos robustos em todos os níveis da organização (ISHIKAWA; ALENCAR, 2020).

Nesse contexto, a classificação de texto surge como uma importante tarefa no campo do Processamento de Linguagem Natural (PLN), uma vez que se pode atribuir rótulos pré-definidos a sequências textuais específicas, auxiliando na identificação e categorização de comportamentos suspeitos.

Historicamente, diversos modelos neurais, como redes convolucionais, redes recorrentes e mecanismos de atenção, foram empregados para o aprendizado de representações textuais (SUN et al., 2020). Com o avanço da inteligência computacional, modelos pré-treinados em extensos corpora, tais como *Global Vectors for Word Representation* (GloVe) (WANG et al., 2019), *Contextualized Word Vectors* (CoVe) (MCCANN et al., 2018) e *Embeddings from Language Models* (ELMo) (PETERS et al., 2018), demonstraram eficácia não apenas na classificação de texto, mas também em várias tarefas de PLN, pavimentando o caminho para o uso de modelos de linguagem de grande porte (LLMs), baseados em arquiteturas de transformadores, como ferramenta promissora na detecção de fraudes em compras públicas. Esses modelos LLM, ao capturarem relações contextuais complexas e oferecerem mecanismos sofisticados de atenção, ampliam as capacidades analíticas, tornando-se um recurso valioso para o desenvolvimento de sistemas robustos e eficientes no combate a fraudes.

2.2 MODELOS DE LINGUAGEM BASEADOS EM TRANSFORMADORES

A introdução de modelos de linguagem baseados em transformadores desencadeou uma revolução no domínio do processamento de linguagem natural. Entre os pioneiros desses modelos estão o GPT (RADFORD et al., 2018) e o BERT (DEVLIN et al., 2018). Estes modelos lograram êxito ao estabelecer um novo paradigma no campo da Inteligência Artificial (IA) de ponta, superando as limitações dos modelos tradicionais de processamento de linguagem natural (WANG et al., 2019).

A premissa fundamental é que os corpora textuais podem ser tão extensos quanto necessário. Assim, o modelo é submetido a um pré-treinamento abrangente para o entendimento da linguagem, valendo-se de um volumoso conjunto de dados textuais. Conforme mencionado por Luccioni, Viguier e Ligozat (2022), o treinamento de tais modelos demanda um investimento significativo de tempo, recursos financeiros e capacidade computacional.

Modelos baseados em transformadores geralmente são pré-treinados em grandes corpora para posterior aplicação em uma variedade de tarefas. Um exemplo disso são os modelos interativos, que são treinados para a previsão da próxima palavra em uma sentença, como o GPT (RADFORD et al., 2018), *Large Language Model Meta AI* (LLaMA) (TOUVRON et al., 2023) e Mistral (JIANG et al., 2023). Existem também os modelos mascarados, assim denominados por serem treinados mediante uma técnica conhecida como *Masked Language Model* (MLM). Este método compreende o mascaramento (substituição) de certas palavras em uma sentença com um token especial, usualmente "[MASK]", sendo o objetivo do modelo prever as palavras mascaradas com base no contexto providenciado pelas palavras não mascaradas na sentença. Exemplos de modelos dessa categoria incluem o BERT (DEVLIN et al., 2018), DeBERTa (HE et al., 2021) e a versão brasileira do BERT, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020).

Ormerod, Patel e Wang (2023) apontam que a vasta maioria dos modelos de linguagem atualmente disponíveis são modelos de palavras mascaradas. Todos esses modelos são variações da mesma arquitetura de transformador apresentada por Vaswani et al. (2017), conforme ilustrado na Figura 2.1.

2.2.1 Arquitetura de um Transformador

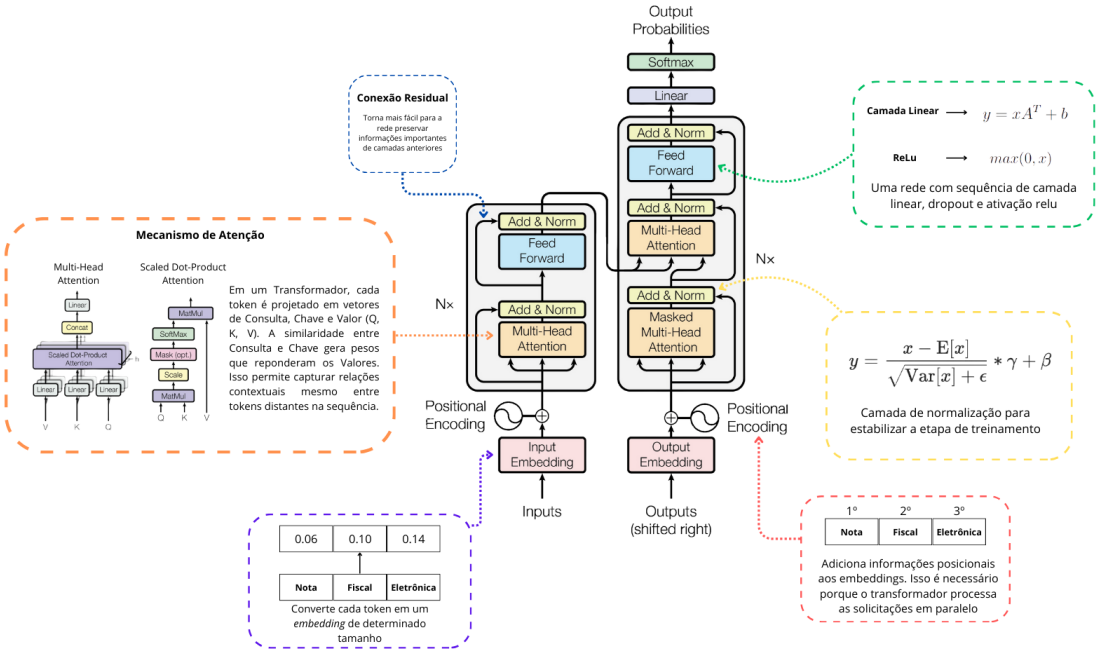
O Transformador, proposto por Vaswani et al. (2017), é uma arquitetura fundamentada em camadas de atenção que dispensam o uso de mecanismos recorrentes ou convolucionais, possibilitando o processamento paralelo de cada token ao longo de toda a sequência. Essa abordagem se mostrou eficiente para lidar com dependências de longo alcance e para acelerar o treinamento, aspectos críticos em tarefas como tradução automática, classificação de texto e sumarização.

O mecanismo de atenção empregado no Transformador avalia a relação entre todos os elementos do texto, atribuindo pesos que indicam a relevância de cada token em relação aos demais. Essa estratégia facilita a identificação de padrões contextuais, reduzindo a necessidade de memorizar estados intermediários, como ocorre em arquiteturas recorrentes, e aprimorando a robustez do modelo frente a sequências extensas ou descrições mais complexas.

A Figura 2.1 apresenta o diagrama completo do Transformador, destacando as duas grandes partes que o compõem: o *encoder* (codificador) e o *decoder* (decodificador). O *encoder* é composto por blocos que combinam atenção e redes *feed-forward*, com conexões residuais que ajudam a evitar problemas de gradiente. No *decoder*, além de um mecanismo de atenção similar, adiciona-se a possibilidade de atender simultaneamente ao contexto gerado no *encoder*, o que é essencial para tarefas de geração de texto ou tradução.

A incorporação posicional (*positional encoding*) e a normalização em camadas (*layer normalization*) fornecem maior estabilidade ao modelo, preservando informações sobre a posição dos tokens e garantindo consistência numérica durante o treinamento. Esse conjunto de técnicas torna o Transformador adaptável a diversas aplicações de Processamento de Linguagem Natural e amplia sua capacidade de lidar com dados heterogêneos e volumosos.

Figura 2.1 Arquitetura completa de um transformador



Fonte: Elaboração própria com dados de Vaswani et al. (2017).

2.2.1.1 *Input Embeddings*

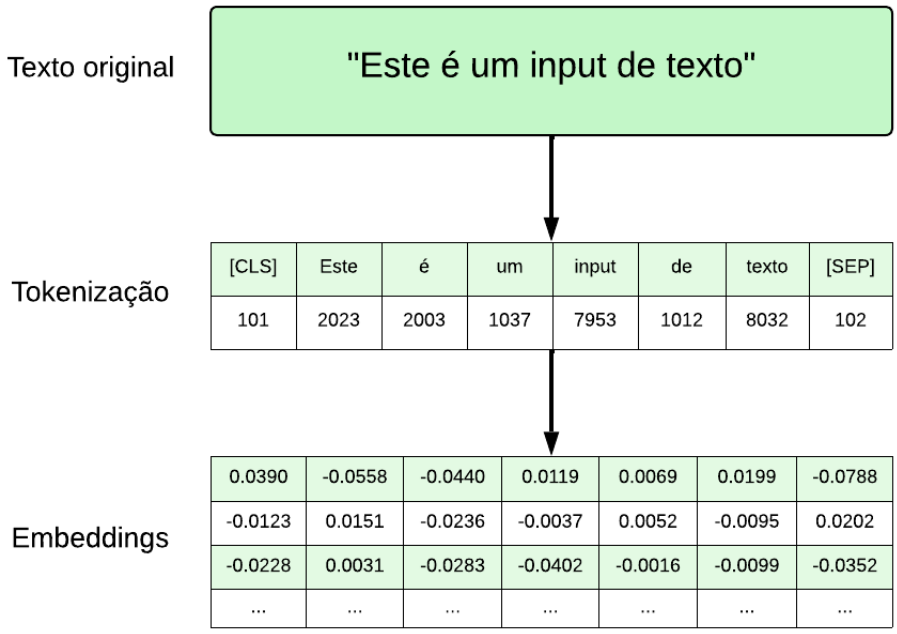
Antes de efetivamente alimentar o modelo com a sequência de tokens, o Transformador (ilustrado na Figura 2.1) precisa converter cada token em uma representação numérica que seja compreensível pelas camadas subsequentes. Para cumprir essa função, o componente de *Input Embeddings* atua como a primeira etapa dos blocos do Transformador, recebendo cada token da sentença em sua forma de ID (índice no vocabulário) e transformando-o em um vetor de N dimensões.

Essa conversão é realizada por meio de uma matriz de *embeddings*, na qual cada linha corresponde a um token específico do vocabulário, e cada coluna, a uma dimensão do vetor de representação. Nesse contexto, o fator $\sqrt{N}$ é aplicado para manter a amplitude numérica dos valores gerados, de modo a preservar a expressividade dos *embeddings* e minimizar problemas de escala durante o treinamento. Dessa forma, ao final do processo de *Input Embeddings*, cada token passa a ser representado por um vetor capaz de capturar nuances semânticas e sintáticas, o que enriquece significativamente a capacidade do modelo de lidar com fenômenos linguísticos complexos.

Conforme a base de dados utilizada varia, as representações geradas podem ser ajustadas para refletir características linguísticas e terminológicas específicas de cada domínio, tornando o modelo mais sensível às variações de contexto. Além disso, a etapa de *Input Embeddings* não atua de forma isolada: ela é complementada por outras técnicas, como a

incorporação posicional (*positional encoding*) e a aplicação de máscaras de atenção, que serão adicionadas em camadas subsequentes. Essas técnicas auxiliam o modelo a levar em conta o ordenamento e a relevância contextual de cada token, de modo que a combinação final de todas as camadas resulte em uma compreensão mais ampla e robusta da sequência textual.

Figura 2.2 *Input Embeddings*



Fonte: Elaboração própria.

2.2.1.2 *Positional Encoding*

Vaswani et al. (2017) adicionam *Positional Encoding* às incorporações de entrada nos blocos codificador e decodificador, fornecendo informações de posição dos tokens. As codificações posicionais e as incorporações têm a mesma dimensão para integração semântica e posicional. No *Positional Encoding*, cria-se uma matriz de codificações posicionais *pe* com dimensões (seq_len, d_model), inicialmente preenchida com 0, aplicando seno aos índices pares e cosseno aos ímpares.

$$\text{Even Indices } (2i): \quad \text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \tag{2.1}$$

$$\text{Odd Indices } (2i+1): \quad \text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \tag{2.2}$$

2.2.1.3 *Layer Normalization*

Nos blocos codificadores e decodificadores, há várias camadas de normalização chamadas **Add & Norm**. O *Layer Normalization* normaliza os dados de entrada calculando sua média e desvio padrão, ajustando para média 0 e desvio padrão 1, com um termo ϵ para evitar divisão por zero. Em seguida, multiplica-se a saída normalizada por α e adiciona-se *bias*, ambos parâmetros ajustáveis durante o treinamento, resultando num tensor normalizado por camadas com escala de entrada consistente para a rede.

2.2.1.4 *Feed-Forward*

A camada de *Feed-Forward* no contexto de um Transformador é uma rede neural totalmente conectada que opera de forma independente em cada posição da sequência de entrada. Estruturalmente, a camada de *Feed-Forward* consiste em duas transformações lineares separadas por uma função de ativação não-linear. A primeira camada linear transforma a dimensionalidade dos dados de entrada, e a função de ativação utilizada é a *Rectified Linear Unit* (ReLU), que introduz não-linearidade no modelo. Em seguida, uma segunda camada linear traz os dados de volta à dimensão original. Matematicamente, essa operação pode ser expressa como:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2.3}$$

Na Equação 2.3, W_1 e W_2 são os pesos, enquanto b_1 e b_2 são os vieses das duas transformações lineares.

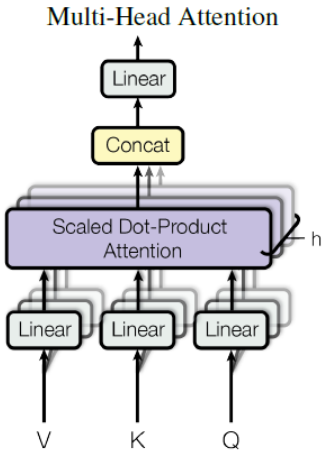
Em síntese, a camada de *Feed-Forward* em cada camada do transformador permite que o modelo desenvolva representações complexas e capture relações não lineares nos dados. Essa capacidade é crítica para o desempenho do modelo, pois amplia sua capacidade de capturar padrões nos valores de entrada. Além disso, como o *Feed-Forward* opera independentemente em cada posição, ele mantém a estrutura posicional dos dados enquanto refina a representação de cada token individualmente.

2.2.1.5 *Multi-Head Attention*

A *Multi-Head Attention* é essencial no Transformador, ajudando o modelo a entender relações complexas nos dados. A Figura 2.3 ilustra o funcionamento da *Multi-Head Attention*.

O bloco *Multi-Head Attention* recebe dados de entrada como consultas, chaves e valores, organizados em matrizes Q, K e V, com dimensões iguais à entrada. Em seguida, cada matriz é transformada linearmente por matrizes de peso $W\hat{Q}$, $W\hat{K}$ e $W\hat{V}$, resultando nas matrizes Q', K' e V'. Estas são divididas em menores, associadas a diferentes *heads* h, permitindo ao modelo processar informações de subespaços diferentes em paralelo. A concatenação de cada *head* em uma matriz H é então transformada por outra matriz de peso W_o, produzindo a saída de *Multi-Head Attention*, a matriz MH-A, preservando a dimensionalidade da entrada.

Figura 2.3 Multi-Head Attention



Fonte: Extraído de Vaswani et al. (2017).

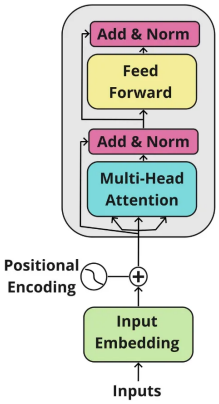
2.2.1.6 *Residual Connection*

Na arquitetura do Transformador, cada subcamada, incluindo a camada de *self-attention* e a *Feed-Forward*, adiciona sua saída à sua entrada antes de passá-la para a camada **Add & Norm**. Esta abordagem integra a saída com a entrada original na camada **Add & Norm**. Este processo é conhecido como **skip connection**, que permite ao Transformador treinar redes profundas de forma mais eficaz, fornecendo um atalho para o gradiente fluir durante a retropropagação.

2.2.1.7 *Encoder*

O *Encoder* mostrado na Figura 2.4 é uma montagem de blocos de codificação que incluem camadas *Multi-Head Attention* e *Feed Forward* com conexões residuais (*Add* e *Norm*). No *Self-Attention*, o mecanismo de atenção usa todas as posições da sequência. O *encoder* gera vetores de consulta, chave e valor, calculando similaridades via produto escalar e normalizando com softmax para obter pesos de atenção. Esses pesos ponderam os valores, refletindo o contexto. Depois da *Self-Attention*, a saída passa por uma rede totalmente conectada com ativação não linear (ReLU), processando cada posição de forma independente para captar representações complexas. Cada subcamada possui normalização de camada e conexão residual para estabilizar o treinamento. O *encoder* transforma uma sequência de vetores de palavra em representações contextuais que incorporam toda a informação da sequência de entrada.

Figura 2.4 Codificador de um Transformador

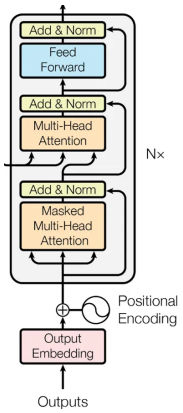


Fonte: Extraído de Vaswani et al. (2017).

2.2.1.8 *Decoder*

Semelhante ao *Encoder*, o *Decoder* (Figura 2.5) consiste em blocos repetidos. A diferença principal é que o *Decoder* tem uma subcamada extra com atenção cruzada, usando a saída do *Encoder* como chaves e valores, e a entrada do *Decoder* como consultas. O *Masked Self-Attention* garante que a previsão em cada posição dependa apenas das posições anteriores, assegurando geração causal do texto. Além disso, cada subcamada no *decoder* inclui normalização de camada e conexões residuais (*Add* e *Norm*). No treinamento, o *decoder* recebe a sequência alvo deslocada, chamada de "teacher forcing", para prever cada token a partir do contexto anterior. O *decoder* integra representações contextuais do *encoder* para captar informações relevantes da entrada. Na inferência, o *decoder* gera a sequência autoregressivamente, começando com um token inicial e construindo a sequência completa sem a sequência alvo real.

Figura 2.5 Decodificador de um Transformador



Fonte: Extraído de Vaswani et al. (2017).

Capítulo 3

TRABALHOS RELACIONADOS

Alguns pesquisadores já se debruçaram sobre a aplicação de modelos de aprendizado de máquina em contratos públicos.

No estudo apresentado por Domingos et al. (2016), os autores visam identificar anomalias nas aquisições na área de Tecnologia da Informação (TI) dentro do Sistema de Aquisições Governamentais do Brasil por meio de técnicas de Deep Learning. O intuito era desenvolver um modelo preditivo destinado a auxiliar na priorização de investigações acerca de compras potencialmente suspeitas, empregando um algoritmo de *Deep Learning* para a análise dos dados de compras de TI realizadas pelo governo federal brasileiro no período de 2014 a 2015. A aplicação da metodologia CRISP-DM juntamente com a ferramenta H2O demonstrou que o modelo final alcançou êxito na detecção de anomalias, dispensando a necessidade de dados previamente etiquetados.

Em Sun e Sales (2018), os autores investigam a aplicação de Redes Neurais Artificiais (RNAs) para a previsão de irregularidades em licitações públicas no Brasil, utilizando dados fornecidos pela Controladoria-Geral da União (CGU). A pesquisa realiza uma comparação do desempenho de dois tipos de RNA, *Deep Neural Network (DNN)* e *Tensor Neural Network (TNN)*, com outros algoritmos, como a regressão logística e a análise de função discriminante. Os resultados demonstraram que a DNN superou significativamente a TNN e os demais algoritmos em diversas métricas de desempenho, sugerindo que a DNN possui maior precisão na predição de irregularidades em licitações públicas.

Lima et al. (2020) utilizam um modelo baseado em uma rede neural *Bidirectional Long Short-Term Memory (Bi-LSTM)* para classificar chamadas de licitação de obras públicas como potencialmente fraudulentas. O estudo emprega modelos de rede neural, especificamente DNN para comparação e Bi-LSTM (modelo principal), utilizando TF-IDF para extração de características. O modelo Bi-LSTM demonstrou alta acurácia na identificação de licitações potencialmente fraudulentas.

Braz et al. (2024) examina as irregularidades em licitações públicas no Brasil, com um foco específico nas pequenas empresas situadas no estado de Minas Gerais. Foram empregadas técnicas de análise exploratória, geoespacial e de redes para a identificação de empresas que apresentem indícios de irregularidades em processos licitatórios. A

análise demonstrou a eficácia dessas abordagens na identificação de pequenas empresas possivelmente envolvidas em atividades fraudulentas, oferecendo *insights* valiosos sobre a interconexão e organização dessas entidades empresariais.

Silva et al. (2024) analisa o problema do sobrepreço em itens de licitações públicas no Brasil, utilizando um conjunto de dados do "Sistema Informatizado de Contas dos Municípios" (SICOM), desenvolvido pelo Tribunal de Contas do Estado de Minas Gerais (TCE-MG). O estudo propõe uma metodologia para processamento e padronização de descrições de itens de licitação e uma abordagem estatística baseada no Intervalo Interquartil (IQR) para detectar sobrepreço, sugerindo que as estratégias avaliadas são promissoras para identificar potenciais irregularidades.

Hott et al. (2023) investigam o uso de modelos baseados em BERT, em especial o BERTopic, para análise de dados de licitações públicas. Os autores utilizaram técnicas de agrupamento de texto e modelagem de tópicos para descobrir padrões ocultos em dados não estruturados. A intenção do estudo foi gerar grupos que capturam os tópicos nos dados de licitação.

Com base nos trabalhos elaborados com dados públicos brasileiros, constata-se que diversas técnicas de aprendizado de máquina, redes neurais e análise estatística foram aplicadas com êxito na identificação de irregularidades em licitações públicas. Os estudos demonstraram que a utilização de algoritmos mais sofisticados, como *auto-encoders* e Bi-LSTM, bem como metodologias estatísticas, contribuem significativamente para a detecção de anomalias em processos de compras governamentais.

No campo internacional, Chen, Huang e Kuo (2009) exploram o uso de redes neurais artificiais (RNAs) para prever litígios relacionados a fraudes em mercados emergentes, especificamente em Taiwan. O objetivo foi aplicar uma rede neural para prever litígios de fraude, auxiliando contadores na elaboração de estratégias de auditoria e melhorando a detecção de fraudes. Os pesquisadores utilizaram redes neurais com topologia de retropropagação para modelar as relações entre variáveis de entrada e saída e prever litígios de fraude, alcançando uma precisão de previsão superior aos métodos tradicionais.

Torres-Berru, Lopez-Batista e Zhingre (2023) exploram a detecção de viés e favoritismo em processos de licitação pública no Equador, utilizando técnicas de mineração de dados e processamento de linguagem natural. Os autores utilizaram um conjunto de dados em espanhol com 1.009.739 perguntas e respostas de 303.076 processos de licitação analisados ao longo de dez anos. Utilizando modelos Word2Vec e o algoritmo VADER, os resultados mostram que houve evidência de favoritismo ou viés de gênero em um percentual significativo dos processos analisados.

Torres-Berru e Batista (2021) examinam o problema da corrupção em licitações públicas no Equador, empregando técnicas de mineração de dados para detectar eventuais irregularidades, como o favorecimento de fornecedores específicos. O objetivo principal consistiu em desenvolver um modelo capaz de identificar padrões anômalos na atribuição de qualificações a contratos públicos de licitação e prever contratos que apresentem anomalias com base nos dados analisados. Para tanto, foi empregado um modelo de múltiplas etapas utilizando K-Means, Mapas Auto-Organizáveis (SOM), *Support Vector Machine* (SVM) e Análise de Componentes Principais (PCA), alcançando alta precisão na detecção de anomalias. Por meio dessa técnica, foram identificados quatro grupos prin-

cipais de processos, dos quais um inclui processos considerados anômalos, caracterizados por atribuir peso insignificante à oferta econômica na decisão final do agente público. Os autores concluem que a aplicação de técnicas de mineração de dados e aprendizado de máquina pode efetivamente auxiliar na identificação e previsão de irregularidades em licitações públicas, promovendo a transparência e reduzindo a corrupção nos processos de contratação pública.

No estudo realizado por Aldana, Falcón-Cortés e Larralde (2022), os autores investigam a aplicação de técnicas de aprendizado de máquina para a identificação e previsão de corrupção em contratos de licitação pública no México. Para tal, empregaram um modelo de *Random Forest* que se revelou eficaz na detecção de contratos considerados anômalos. De acordo com os autores, as variáveis preditoras de maior relevância foram aquelas associadas ao relacionamento entre compradores e fornecedores, tais como o valor total despendido por um comprador junto a um fornecedor e a proporção de contratos de licitação única concedidos a um determinado fornecedor.

Nai, Sulis e Meo (2022) apresentam uma revisão sistemática sobre a detecção de fraudes em licitações públicas usando técnicas de inteligência artificial (IA), incluindo aprendizado de máquina, redes neurais e processamento de linguagem natural, destacando a eficácia dessas técnicas na melhoria da capacidade de detecção de fraudes em processos de licitação pública. Os resultados revelaram que diversas técnicas de IA são aplicadas na detecção de fraudes em licitações públicas, incluindo algoritmos de classificação, redes neurais e análise de redes sociais. As disciplinas mais envolvidas incluem Ciência da Computação, Engenharia e Gestão de Negócios. Os estudos utilizam predominantemente dados de licitações públicas para treinar modelos preditivos.

Costa et al. (2022) discutem a identificação de fraudes em licitações públicas através da aplicação combinada de trilhas de auditagem (sequências de regras voltadas à detecção de possíveis irregularidades) e redes sociais. O propósito da investigação foi desenvolver uma metodologia que permitisse identificar indícios de fraude, com os dados fornecidos pelo Ministério Público de Minas Gerais (MPMG). Os pesquisadores desenvolveram um modelo para analisar as relações entre empresas licitantes e seus sócios, utilizando dados como sócios, e-mails, telefones e endereços comuns. Os resultados indicam que essa abordagem pode reduzir a quantidade de documentos que precisam de análise manual, tornando o combate à corrupção mais eficiente. No entanto, o estudo enfrenta limitações devido à falta de dados históricos completos, o que pode levar a falsos positivos. Futuramente, os pesquisadores pretendem implementar auditorias adicionais, como a identificação de sócios de empresas que também ocupam cargos públicos, uma situação frequentemente ligada a fraudes.

Em suma, os estudos revisados evidenciam que várias abordagens de IA e técnicas de mineração de dados, têm se mostrado eficazes na identificação de irregularidades em processos de licitação pública, tanto no Brasil como em contextos internacionais. Estas metodologias, quando aplicadas de forma diversificada, constituem ferramentas valiosas para o fortalecimento dos mecanismos de controle e auditoria nas contratações governamentais.

No entanto, embora eficazes, a maioria das abordagens anteriormente discutidas exige a rotulagem prévia de dados por especialistas na área de fraudes para possibilitar o trei-

namento. Ademais, as bases de dados empregadas são majoritariamente compostas por informações históricas, caracterizadas por um elevado atraso temporal, o que pode dificultar a análise de casos recentes e resultar em lentidão na atualização dos modelos para novos problemas emergentes. Ademais, é importante destacar que nenhuma dessas abordagens explorou a aplicação de modelos de aprendizado de máquina diretamente sobre dados de compras, tais como descrições de produtos, valores, entre outros. Neste contexto, o presente estudo complementa as investigações anteriores ao explorar diretamente dados atualizados das compras governamentais brasileiras e aplicar modelos de PLN fundamentados em *Large Language Model* (LLM), juntamente com técnicas como *zero-shot*, *few-shot* e *fine-tuning*, o que permite minimizar a necessidade de rotulagem manual intensiva e aprimorar a abrangência e eficácia na detecção automática de empresas potencialmente inidôneas.

Capítulo 4

DETECÇÃO DE EMPRESAS INIDÔNEAS EM COMPRAS GOVERNAMENTAIS

4.1 RECURSOS COMPUTACIONAIS E CÓDIGOS

A coleta e o tratamento dos dados utilizados nesta pesquisa foram realizados em uma máquina local, ao passo que o treinamento dos modelos foi realizados no ambiente Google Colab. A Tabela 4.1 sintetiza a infraestrutura empregada nas diferentes etapas da pesquisa. Em termos de capacidade de memória, a máquina local mostrou-se suficiente para as tarefas de limpeza, tratamento e serialização dos conjuntos em formato Parquet, garantindo baixo overhead de leitura/escrita e compatibilidade direta com as rotinas implementadas em Python. Já o Colab, ofereceu os recursos necessários para executar, de forma reproduzível, experimentos nas configurações *zero-shot*, *few-shot* e *fine-tuning*, particularmente vantajosos para a etapa de otimização de hiperparâmetros e ajuste fino de grandes modelos de linguagem. A versão do Python utilizada na maior parte do trabalho foi a 3.11.2.

Tabela 4.1 Recursos computacionais utilizados

	Máquina Local	**Google Colab**
RAM do Sistema	20 GB	83,5 GB
Disco	2 TB	235,7 GB
Placa de Vídeo	–	NVIDIA A100
RAM da GPU	–	40 GB

Fonte: Elaboração própria.

Os códigos desenvolvidos durante este trabalho de pesquisa encontram-se disponíveis em "https://github.com/CleitonOERocha/Mestrado", estruturados em jupyter notebooks e códigos ".py".

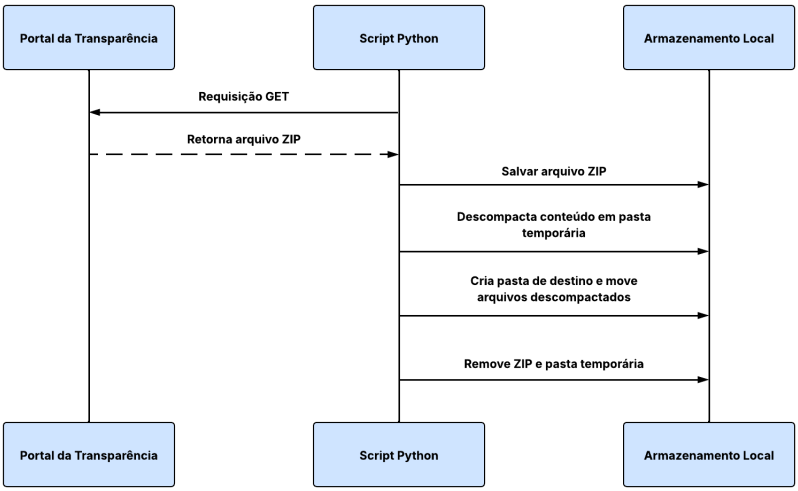
Os datasets tratado para os anos de 2022, 2023 e 2024 encontram-se disponíveis em "https://drive.google.com/drive/folders/17RTCP2wF52In7lIQcwhyOEmw5FG8Yleu" no formato Parquet.

4.2 COLETA E TRATAMENTO DOS DADOS

O diagrama de sequência apresentado na figura 4.1 ilustra o fluxo de interação relacionado à coleta e armazenamento de notas fiscais eletrônicas (NF-e). Os componentes envolvidos são o site do Portal da Transparência, o código em Python e o armazenamento local. Para a coleta dos arquivos, utilizou-se o link base do site do Portal da Transparência, cuja URL é associada ao ano e mês de referência dos arquivos:

"https://dadosabertos-download.cgu.gov.br/PortalDaTransparencia/saida/nfe/" + **ano** + **mês** + "_NFe.zip"

Figura 4.1 Requisição GET para extração web



Fonte: Elaboração própria.

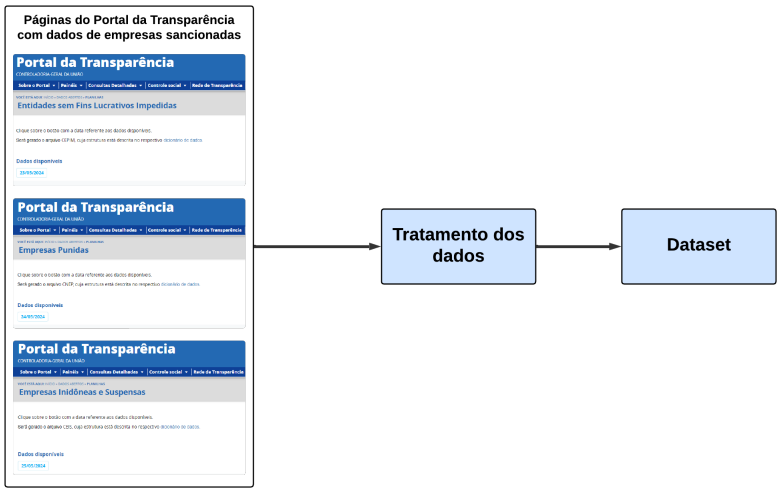
Inicialmente, a interação começa com o código Python efetuando uma requisição do tipo GET para o Portal da Transparência, solicitando especificamente dados referentes às NF-e. Segundo Alam, Cartledge e Nelson (2014), a requisição GET é um método do protocolo HTTP cuja função principal é obter (ou "recuperar") a representação de um recurso em um servidor, sem produzir alterações no estado desse recurso. Dessa forma, quando um cliente faz uma requisição GET para determinado endereço, espera-se que o servidor retorne algum tipo de resposta, geralmente com um corpo de dados (por exemplo, uma página HTML), sem que essa chamada modifique o que já está armazenado ou publicado ali. Em resposta a essa requisição, o Portal da Transparência retorna os dados solicitados para o código Python, permitindo assim o processamento dos dados recebidos.

Uma vez obtidos esses dados, o código Python realiza a etapa seguinte do processo, que é a persistência das informações coletadas. Esse passo é feito através da operação de armazenamento dos dados localmente. Com isso, conclui-se o ciclo da interação, marcado pela obtenção inicial dos dados do Portal, tratamento por meio do código Python e subsequente armazenamento local das informações coletadas.

Posteriormente, os arquivos coletados são submetidos a um processo de tratamento, que abrange a limpeza, a padronização de formato e a extração das informações pertinentes.

Para realizar a coleta dos dados das empresas punidas na esfera pública foram utilizados os dados provenientes da GCU e disponíveis no Portal da Transparência[1][2][3]. O cadastro é composto por empresas punidas e/ou suspensas de participar de licitações. Na Figura 4.2 é demonstrada o fluxo de obtenção dessas empresas. Após a coleta, essas informações passam por um tratamento focado em padronizar colunas, remover duplicidades e adequar os campos de identificação, resultando em uma lista de empresas sancionadas. O formato de coleta dos dados na *web* segue o mesmo rito dos dados das notas fiscais apresentado na Figura 4.1.

Figura 4.2 Fluxograma de coleta de informações das empresas



Fonte: Elaboração própria.

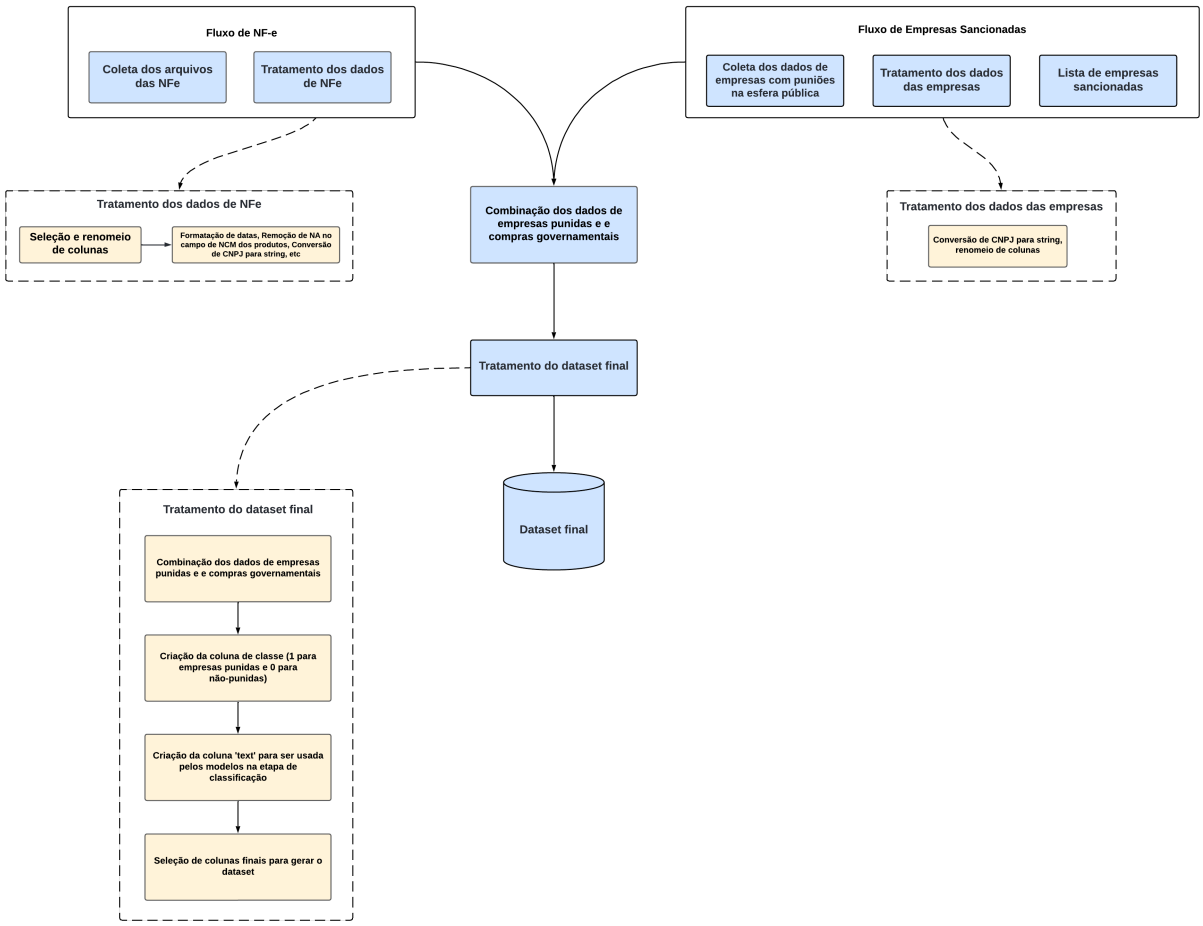
O fluxograma ilustrado na Figura 4.3 apresenta o processo geral para formação do dataset final. Na seção superior (Fluxo de NF-e), encontra-se a etapa de coleta dos arquivos das Notas Fiscais Eletrônicas (NF-e). Já na parte inferior (Fluxo de Empresas Sancionadas), são coletados os dados de companhias que receberam punições na esfera pública. Em seguida, ambas as frentes convergem na etapa de "Unificar Dados", onde as informações processadas das NF-e são cruzadas com a lista de empresas sancionadas. Esse passo de unificação, conduzido em Python, envolve a combinação de chaves (como CNPJ) para relacionar as Notas Fiscais Eletrônicas às empresas que sofreram sanções, ampliando o escopo analítico dos registros.

Por fim, o resultado do *merge* forma o "Dataset final", onde se consolida todo o histórico de operações e a identificação das empresas envolvidas. Esse dataset serve de

[1]https://portaldatransparencia.gov.br/download-de dados/ceis
[2]https://portaldatransparencia.gov.br/download-de-dados/cnep
[3]https://portaldatransparencia.gov.br/download-de-dados/cepim

base para análises aprofundadas, verificação de padrões de comportamento e potencial detecção de irregularidades.

Figura 4.3 Fluxograma geral para construção do dataset final



Fonte: Elaboração própria.

O processo de tratamento dos dados se inicia com a leitura dos arquivos de Notas Fiscais Eletrônicas (NF-e), referentes a um determinado ano, a partir de um diretório onde os arquivos brutos são armazenados. Cada arquivo é aberto e passa por uma seleção de colunas para manter apenas aquelas de interesse para a análise, seguida do renomeio dessas colunas para padrões mais consistentes. Em seguida, remove-se qualquer registro cujos campos de classificação do produto estejam ausentes, além de serem convertidos os formatos de dados, como datas e campos de CNPJ, ajustando-se, por exemplo, o preenchimento de zeros à esquerda quando necessário. Todos esses arquivos são então unificados em um único dataframe, consolidando as informações relativas às transações do ano especificado.

Após a etapa de preparação dos dados das NF-e, realiza-se a integração com a base

de empresas punidas, mostrada no lado esquerdo da figura, que envolve a leitura de uma lista de companhias sujeitas a sanções públicas. Nesse momento, o CNPJ dessas empresas é também tratado para manter consistência no formato, de modo que a junção entre as duas bases seja possível pela chave de CNPJ do emitente. A junção resulta na criação de uma coluna denominada "classe", que recebe o valor "1" para empresas punidas e "0" para as demais. Além disso, é construída uma coluna de texto ("text") que agrega diversos campos das notas fiscais em um único texto, atendendo aos requisitos de modelos de classificação baseados em linguagem natural. Por fim, faz-se a seleção final das colunas de interesse e gera-se um arquivo unificado no formato Parquet.

4.3 MODELOS

Os modelos utilizados neste trabalho estão disponíveis na Tabela 4.2. Ao todo, foram avaliados 14 modelos, abrangendo diversas configurações e tamanhos, selecionados a partir dos grupos BERT, Deepseek, Gemma, GPT, LLaMA e Mistral. Esses grupos foram escolhidos pela sua relevância atual, capacidade técnica comprovada e pelo destaque recente na comunidade científica e tecnológica.

A seleção dos modelos foi orientada pelos seguintes critérios: a compatibilidade com a unidade de processamento gráfico NVIDIA A100 (utilizada no Google Colab); a data de lançamento, favorecendo modelos mais recentes em virtude do progresso contínuo da tecnologia; a habilidade dos modelos em executar tarefas de classificação textual, um elemento essencial para a aplicação proposta neste estudo; e, por último, a disponibilidade para uso gratuito, sem a restrição ao uso de APIs pagas.

Tabela 4.2 Modelos utilizados

Grupo	**Modelo**
Bert	neuralmind/bert-base-portuguese-cased
Bert	neuralmind/bert-large-portuguese-cased
Deepseek	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
Deepseek	deepseek-ai/DeepSeek-R1-Distill-Llama-8B
Gemma	google/gemma-3-12b-pt
Gemma	google/gemma-2-2b
Gemma	google/gemma-2b-it
GPT	openai-community/gpt2
GPT	openai-community/gpt2-large
LLaMA	meta-llama/Llama-3.2-3B
LLaMA	meta-llama/Llama-3.1-8B
LLaMA	meta-llama/Llama-2-7b-hf
Mistral	mistralai/Mistral-7B-v0.1
Mistral	mistralai/Mistral-7B-Instruct-v0.3

Fonte: Elaboração própria.

Todos os modelos utilizados estão disponíveis publicamente na plataforma Hugging

Face (WOLF et al., 2020), facilitando a replicação do estudo e permitindo comparações futuras com novos modelos ou atualizações das versões atuais. As seções subsequentes detalham ainda mais cada grupo de modelos, abordando suas particularidades técnicas e apresentando comparações de desempenho obtidas nos experimentos realizados.

4.3.1 LLaMA

Touvron et al. (2023) apresenta a família de modelos LLaMA, composta por redes de 7 a 65 bilhões de parâmetros treinadas exclusivamente com dados públicos e destinados a oferecer desempenho de ponta com menor custo de inferência. Para construir esses modelos, a equipe seguiu as leis de escalonamento de Chinchilla (HOFFMANN et al., 2022), porém preferiu aumentar significativamente o número de tokens em vez de inflar o número de parâmetros. O pré-treino abrangeu cerca de 1,4 trilhão de tokens, provenientes de uma mistura diversa de fontes públicas: *CommonCrawl* processado com CCNet, C4, repositórios GitHub sob licenças permissivas, Wikipédia, livros de domínio público (Gutenberg e Books3), artigos do arXiv e dados do *Stack Exchange*. O *CommonCrawl* representou dois terços do corpus, enquanto cada uma das demais fontes contribui de 2 a 15%.

A arquitetura partiu do *Transformer*, mas incorpora três modificações principais. Primeiro, aplica normalização *pre-Layer* com RMSNorm, inspirada no GPT-3, para maior estabilidade. Segundo, substitui a ReLU pela ativação SwiGLU, reduzindo o tamanho da camada *feed-forward* para dois terços do habitual. Terceiro, eliminou *embeddings* absolutos e insere *Rotary Positional Embeddings* (RoPE) em todas as camadas. Essas alterações visaram melhorar a eficiência de treino e de inferência sem sacrificar qualidade.

Tabela 4.3 Arquitetura do LLaMA

params	***dimension***	n_{heads}	n_{layers}	***learning rate***	***batch size***	n_{tokens}
6.7B	4.096	32	32	3.0×10^{-4}	4M	1.0T
13.0B	5.120	40	40	3.0×10^{-4}	4M	1.0T
32.5B	6.656	52	60	1.5×10^{-4}	4M	1.4T
65.2B	8.192	64	80	1.5×10^{-4}	4M	1.4T

Fonte: Extraído de Touvron et al. (2023).

Quatro tamanhos foram disponibilizados. O modelo de 6,7 B usa dimensão de estado 4.096 com 32 camadas; o de 13B, 5.120 com 40 camadas; o de 32,5B, 6.656 com 60 camadas; e o de 65,2 B, 8.192 com 80 camadas. Todos empregam 32 a 64 *multi-head attention* e um lote global de 4 milhões de tokens. Os dois menores foram treinados em 1T token; os maiores, em 1,4T tokens, preservando a mesma proporção de dados.

O treinamento utilizado foi o AdamW com $\beta_1 = 0,9$, $\beta_2 = 0,95$ e decaimento de peso igual a 0,1. Para conter memória e acelerar cálculos, os autores adotaram atenção otimizada da biblioteca xFormers, *checkpointing* seletivo e paralelismo de modelo e de sequência. O LLaMA-65B processa cerca de 380 tokens por segundo e levou cerca de 21 dias para treinar, utilizando 2.048 GPUs A100-80 GB.

Nas avaliações, o LLaMA-13B, dez vezes menor que o GPT-3, apresentou altas taxas nos *benchmarks* de raciocínio e compreensão, enquanto o LLaMA-65B se manteve no nível de modelos muito maiores como PaLM-540B. Esse desempenho ocorreu tanto da maior quantidade de dados por parâmetro quanto das otimizações de arquitetura e de treinamento.

4.3.2 Gemma

Introduzida em Team et al. (2024), a família Gemma, constituída por modelos de 2 e 7 bilhões de parâmetros, foi desenvolvida a partir da pesquisa e infraestrutura que deram origem à linha Gemini, mas com pesos disponibilizados de forma aberta.

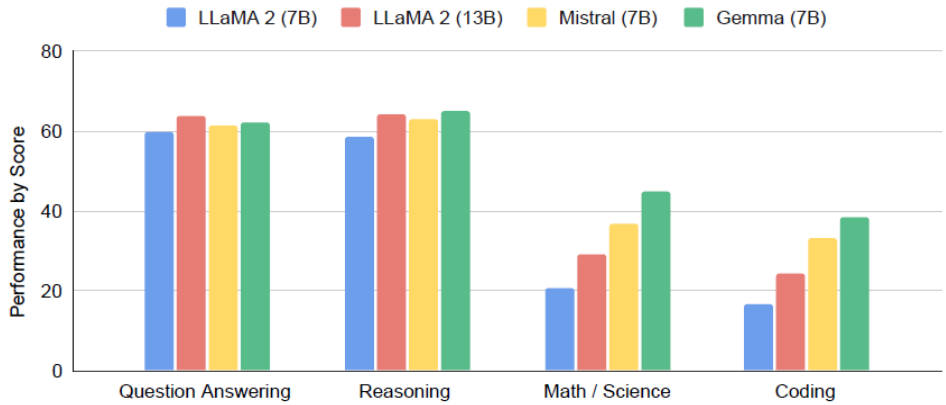
A Gemma foi construída sobre um decodificador *Transformer* treinado para contextos de até 8.192 tokens. Sua arquitetura mantém o mecanismo de *multi-head attention* na versão de 7B, enquanto adota atenção multi-query no modelo de 2B, uma escolha que Team et al. (2024) consideraram mais eficiente para essa escala. Em ambas as variantes, são implementadas melhorias pós-Transformer, como o RoPE, ativações GeGLU e normalização RMSNorm, além de compartilhamento de *embeddings* para otimizar o tamanho do modelo.

Tabela 4.4 Arquitetura do Gemma

params	$d_{\textbf{model}}$	n_{layers}	**FF dim**	n_{heads}	n_{KV}	***head size***	***vocab size***
2B	2.048	18	32.768	8	1	256	256.128
7B	3.072	28	49.152	16	16	256	256.128

Fonte: Extraído de Team et al. (2024).

Figura 4.4 Comparativo entre Gemma e outros modelos em diferentes tarefas



Fonte: Extraído de Team et al. (2024).

A fase de pré-treinamento empregou 3 trilhões de tokens para a versão de 2B e 6 trilhões para a de 7B, utilizando dados predominantemente em inglês oriundos da Web.

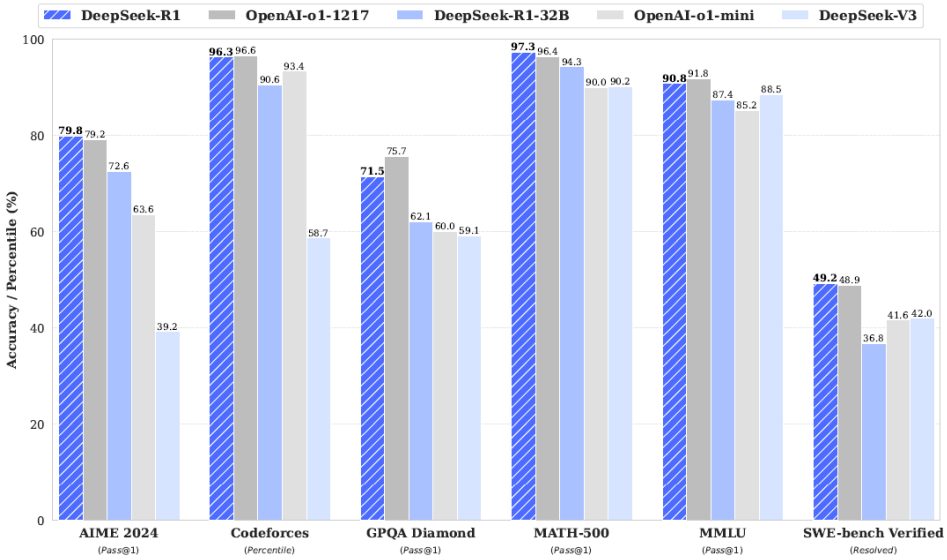
Após o pré-treinamento, cada versão foi submetida a um processo de *fine-tuning* supervisionado com pares de instrução e resposta, tanto sintéticos quanto humanos, seguido de reforço por meio de feedback humano (RLHF). O processo de formatação incluiu tokens de controle para delimitar o início e o término de cada turno.

Nos testes, a Gemma 7B superou LLaMA 2 13B e Mistral 7B em métricas como MMLU, GSM8K e HumanEval, demonstrando uma vantagem notável em matemática e código; a versão de 2B também excedeu o desempenho típico de modelos de tamanho comparável. Avaliações humanas de aproximadamente 1.000 prompts atribuíram ao Gemma 7B um índice de vitória de 61% em tarefas de instrução e 63% em segurança, em comparação com o Mistral 7B Instruct.

4.3.3 *DeepSeek*

Apresentado inicialmente em DeepSeek-AI et al. (2025), o DeepSeek-R1, primeira geração de modelos de raciocínio da equipe DeepSeek adota uma arquitetura *Mixture-of-Experts* (MoE), em que apenas uma parte dos parâmetros fica ativa em cada inferência, são 37 bilhões de parâmetros efetivos dentro de um total de 671 bilhões. Segundo DeepSeek-AI et al. (2025), essa escolha buscou conciliar capacidade expressiva e eficiência computacional.

Figura 4.5 Comparativo entre DeepSeek e outros modelos em diferentes tarefas



Fonte: Extraído de DeepSeek-AI et al. (2025).

O processo de treinamento foi organizado em quatro etapas. Primeiro, coletou-se "dados de arranque", milhares de cadeias de raciocínio longas e legíveis, para realizar um *fine-tuning* no modelo DeepSeek-V3-Base, de modo a obter um ponto inicial estável para o reforço. Em seguida, realizou-se um estágio de reforço orientado a raciocínio sobre tarefas de matemática, código, ciências e lógica, incorporando uma recompensa extra que penalizava a mistura indesejada de idiomas. Quando esse reforço convergiu, os pesquisadores aplicaram uma técnica chamada *rejection sampling* sobre o próprio modelo

para gerar um conjunto supervisionado: cerca de 600 mil exemplos de raciocínio e 200 mil de tarefas gerais, totalizando 800 mil amostras. Depois de duas épocas de *fine-tuning* com esse material, um segundo estágio de reforço integrou todos os cenários, equilibrando utilidade e segurança, até produzir o *checkpoint* denominado DeepSeek-R1. Por fim, o resultado desse treinamento é transferido para modelos menores baseados em Qwen e Llama usando o mesmo conjunto de 800 mil exemplos.

Nos testes, o DeepSeek-R1 alcançou 79,8% de acerto na primeira tentativa no AIME 2024, 97,3% no conjunto MATH-500 e 90,8% no MMLU; em competições de código, obteve nota de especialista, superando 96,3% dos participantes do Codeforces. Esses resultados o colocaram no mesmo patamar do modelo OpenAI-o1-1217, superando modelos abertos anteriores em tarefas de raciocínio, redação criativa e compreensão de contexto longo.

A estratégia empregada no desenvolvimento do DeepSeek-R1 demonstrou que grandes modelos podem ser desenvolvidos quase inteiramente por reforço, enquanto a destilação transfere padrões de pensamento para redes mais compactas com alto ganho de desempenho. Assim, o trabalho contribuiu tanto com avanços metodológicos, por exemplo, com um pipeline de duas etapas de SFT e duas de RL, quanto com modelos abertos que elevam o estado da arte em *benchmarks* de matemática, programação e conhecimento geral.

4.3.4 *BERT*

Souza, Nogueira e Lotufo (2020) apresentam o BERTimbau, uma versão monolíngue do BERT para o português brasileiro. O BERTimbau foi construído em duas versões – Base e Large – que mantêm a arquitetura original do *Transformer Encoder*, mas diferem em escala: a variante Base possui 12 camadas, dimensão oculta de 768, 12 *multi head attention* e cerca de 110 milhões de parâmetros, enquanto a Large eleva esses números para 24 camadas, 1024 unidades, 16 *multi head attention* e aproximadamente 330 milhões de parâmetros.

Para adequar o modelo à língua, Souza, Nogueira e Lotufo (2020) construíram um vocabulário, obtido com *SentencePiece* e posteriormente convertido ao formato *WordPiece*, um método que divide palavras em segmentos menores quando necessário, reduzindo casos fora de vocabulário e encurtando sequências tokenizadas. O treinamento foi realizado a partir de *checkpoints* de modelos BERT já existentes (mBERT para a versão Base, BERT-Large em inglês para a versão Large) e prosseguiu utilizando as tarefas autosupervisionadas de modelagem de palavras mascaradas e previsão da próxima sentença. A fase Base usou sempre sequências de 512 tokens, enquanto a Large começou com 128 tokens e, nos 100 mil passos finais, adota sequência completa de 512 tokens.

O corpus de pré-treinamento foi o *Brazilian Web as Corpus* (brWaC), um arranjo de páginas da web brasileiras que reúne 2,68 bilhões de tokens em 3,53 milhões de documentos. Após limpeza de código HTML e correção de caracteres, restaram cerca de 17,5 GB de texto bruto que alimentaram o processo de geração de exemplos para as tarefas de pré-treino, permitindo cobertura lexical ampla e diversidade temática.

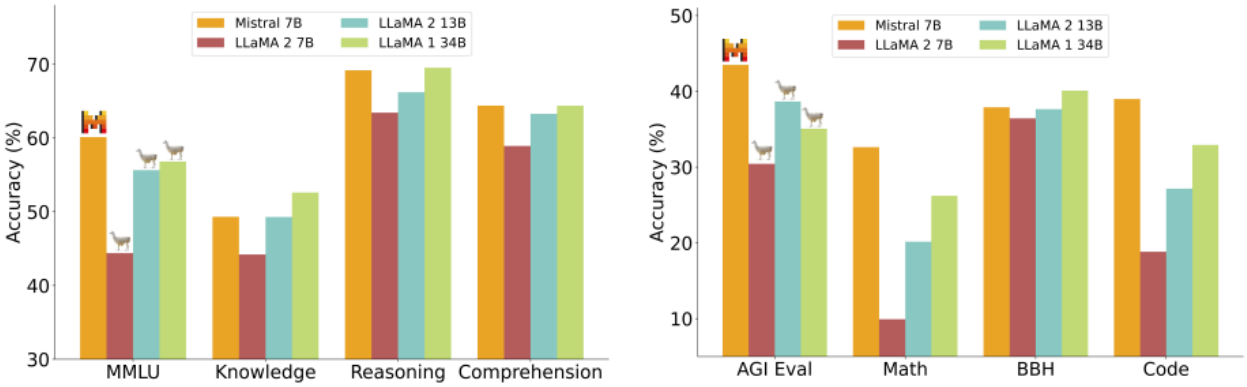
Os modelos foram avaliados em três tarefas de PLN. Na ASSIN2, que reúne Simi-

laridade Textual de Sentença (STS) e Reconhecimento de Inferência Textual (RTE), o BERTimbau-Large obteve correlação de Pearson de 0,852 para STS e F1 de 90% para RTE, superando sistemas anteriores que recorriam a *ensembles* ou a modelos traduzidos para o inglês; a versão Base também ultrapassou o mBERT com *fine-tuning*. Na tarefa de reconhecimento de entidades nomeadas, o BERTimbau-Large atingiu F1 de 78,5% no cenário total, frente a 73,1% do mBERT equivalente.

4.3.5 *Mistral*

O Mistral (JIANG et al., 2023) é um modelo de linguagem baseado na arquitetura *Transformer* com 7 bilhões de parâmetros, desenvolvido com o objetivo principal de alcançar alto desempenho em tarefas diversas mantendo eficiência computacional e tempo reduzido de inferência. Destaca-se por empregar duas modificações importantes em relação a arquiteturas anteriores como a do Llama: o *Grouped-query Attention* (GQA) e o *Sliding Window Attention* (SWA). O GQA melhora significativamente a velocidade da inferência ao reduzir os requisitos de memória durante a decodificação, permitindo maiores tamanhos de lotes, enquanto o SWA limita a atenção a uma janela fixa de tokens (no caso, 4.096 tokens), permitindo maior eficiência para lidar com sequências muito longas. Segundo Jiang et al. (2023), isso possibilita que o modelo tenha uma atenção teórica de até cerca de 131 mil *tokens* em suas 32 camadas, garantindo desempenho competitivo mesmo com sequências acima de 16 mil *tokens*. Por fim, uma inovação adicional introduzida pelo Mistral 7B é o uso de um mecanismo chamado *Rolling Buffer Cache*, que permite limitar a memória utilizada durante a inferência ao sobrescrever periodicamente dados antigos de atenção, resultando em redução do consumo de memória de até 8 vezes sem perda significativa de qualidade.

Figura 4.6 Comparativo entre Mistral e outros modelos em diferentes tarefas



Fonte: Extraído de DeepSeek-AI et al. (2025).

Nos ensaios de avaliação mencionados por Jiang et al. (2023), o Mistral 7B consistentemente supera outros modelos, tais como Llama 2 e Llama 1, em diversos critérios relacionados ao raciocínio lógico, matemática, geração de código e compreensão geral (Fi-

gura 4.6). Particularmente, o modelo exibe um desempenho notável nos testes GSM8K e MATH em relação à matemática, bem como no Humaneval e MBPP no tocante à geração de código, chegando próximo ao rendimento de modelos especializados em código, como o Code-Llama. Tal evidência destaca a competência do Mistral 7B em integrar diferentes áreas do conhecimento, mantendo alta eficiência computacional e demonstrando-se superior aos modelos Llama em todas as categorias avaliadas. Ademais, foi criada e testada uma versão do modelo, denominada Mistral 7B-Instruct, projetada especificamente para seguir instruções. Esta versão demonstrou um desempenho superior em testes de benchmarks interativos (MT-Bench), alcançando resultados comparáveis ou superiores aos do modelo Llama 2 Chat, que possui 13 bilhões de parâmetros.

4.4 MÉTRICAS DE DESEMPENHO

Para tarefas de classificação binária, a forma mais tradicional de mensurar o desempenho dos modelos de aprendizado supervisionado geralmente é a partir da matriz de confusão (SKIENA, 2017). Essa matriz é composta por quatro categorias de resultados: verdadeiro positivo (TP), verdadeiro negativo (TN), falso positivo (FP) e falso negativo (FN). Os termos "verdadeiro" e "falso" referem-se ao resultado da predição do modelo, enquanto "positivo" e "negativo" dizem respeito às classificações reais. Em essência, a matriz de confusão confronta as predições do modelo com as respostas corretas, indicando se o modelo acertou ou errou, e qual foi a natureza desses acertos e erros.

Tabela 4.5 Matriz de confusão para problemas de duas classes

	Predição Positiva	**Predição Negativa**
Classe Positiva	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: Elaboração própria.

A partir da matriz de confusão, é possível calcular diversas métricas de desempenho, como acurácia, precisão, cobertura (*recall*) e F1-Score (SKIENA, 2017). A acurácia mede o desempenho geral do modelo, mas não considera possíveis desbalanceamentos nos dados, como quando uma classe é muito mais frequente que outra. Já a precisão avalia a proporção de acertos entre os casos preditos como positivos pelo modelo. A cobertura, por sua vez, mede a proporção de acertos em relação aos casos positivos reais. Por fim, o F1-Score calcula a média harmônica entre a precisão e a cobertura, produzindo um valor entre 0 e 1, onde valores mais próximos de 1 indicam melhores resultados.

Outra métrica amplamente utilizada é a curva ROC (*Receiver Operating Characteristic*) (PARK; GOO; JO, 2004). Ela relaciona a taxa de verdadeiros positivos (TPR) com a taxa de falsos positivos (FPR) para diferentes limiares de classificação. Para avaliar o modelo com uma métrica única a partir dessa curva, utiliza-se a área abaixo da curva, conhecida como AUC (*Area Under the Curve*). A AUC reflete a capacidade geral do modelo em distinguir entre as classes e é especialmente útil em cenários com desbalanceamento entre as classes.

Embora a matriz de confusão e as métricas derivadas sejam amplamente utilizadas, elas apresentam limitações em cenários nos quais os custos associados a diferentes tipos de erro de classificação variam significativamente (MONARD; BARANAUSKAS, 2003). Nesse contexto, conforme indicado por Prati, Batista e Monard (2003), a matriz de custo emerge como uma ferramenta complementar, atribuindo penalidades distintas para erros de classificação, permitindo ao modelo priorizar a minimização dos custos totais em vez de apenas otimizar métricas como a acurácia. Por exemplo, classificar um caso positivo como negativo (falso negativo) pode ter um impacto mais severo do que classificar um caso negativo como positivo (falso positivo), dependendo da aplicação. Portanto, a matriz de custo ajusta o processo de treinamento para considerar essas diferenças.

Tabela 4.6 Métricas de desempenho

Acurácia	**Precisão**
$\frac{TP+TN}{TP+TN+FP+FN}$ (4.1)	$\frac{TP}{TP+FP}$ (4.2)
Recall	**F1-Score**
$\frac{TP}{TP+FN}$ (4.3)	$2 \times \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}}$ (4.4)

Fonte: Elaboração própria.

Ainda nessa linha, Prati, Batista e Monard (2003) indica que a matriz de custo pode ser integrada à matriz de confusão para avaliar o desempenho do modelo de maneira ponderada. Em vez de considerar apenas o número de acertos e erros, ela incorpora os custos específicos associados a cada tipo de erro, convertendo a análise tradicional em uma abordagem mais alinhada aos objetivos práticos do sistema. Essa integração se mostra particularmente relevante em aplicações como detecção de fraudes ou diagnóstico médico, onde o impacto de diferentes tipos de erro pode variar drasticamente.

A matriz de custo neste trabalho foi construída conforme demonstrada por Wang (2018), utilizando zeros para os resultados Verdadeiro-Positivo e Verdadeiro-Negativo para não penalizar quando o modelo acerta, enquanto os valores Falso-Positivo e Falso-Negativo sofrem uma penalização por confundirem uma classe com a outra. A matriz é então exposta como $\begin{bmatrix} 0 & 5 \\ 1 & 0 \end{bmatrix}$, sendo construída para um problema binário em que rotular uma compra governamental como potencialmente realizada por uma empresa sancionada é considerado cinco vezes pior do que o erro inverso (FP = 5/FN = 1); para fins de comparação e análises, também foi criada uma matriz com peso cinco para os rótulos registrados como negativo quando o correto era positivo (FN = 5/FP = 1). Para chegar a um valor final, a fórmula é expressa por:

$$\text{Custo FP} = 5 \text{ / FN} = 1 : \frac{(\text{Qtd. Registros FP} \times 5) + \text{Qtd. Registros FN}}{\text{Total de Registros}} \tag{4.1}$$

$$\text{Custo FN} = 5 \text{ / FP} = 1 : \frac{(\text{Qtd. Registros FN} \times 5) + \text{Qtd. Registros FP}}{\text{Total de Registros}} \tag{4.2}$$

Dessa forma, torna-se possível comparar as técnicas e modelos empregados, mesmo com uma diferença entre o número de registros usados na etapa de validação.

Por fim, a análise ROC também pode complementar o uso de matrizes de custo, pois possibilita a visualização de como o modelo se comporta em diversos cenários de custos e distribuições de classes. A área sob a curva ROC (AUC) reflete o desempenho do modelo sob uma perspectiva global, enquanto a matriz de custo oferece insights específicos sobre as implicações práticas dos erros de classificação. Juntas, essas ferramentas oferecem uma visão abrangente e prática para a avaliação e aprimoramento dos modelos avaliados.

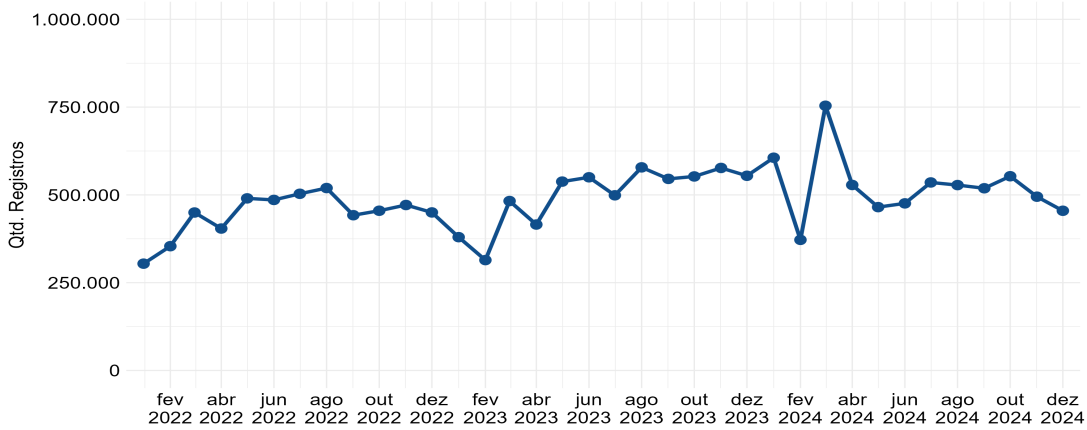
Capítulo 5

RESULTADOS

5.1 ANÁLISE EXPLORATÓRIA

A análise exploratória dos dados das notas fiscais identificou padrões temporais e espaciais nos registros, além de características associadas à classificação de empresas em categorias de punidas e não punidas. O total de registros mensais, conforme ilustrado na Figura 5.1, demonstra flutuações ao longo do período de fevereiro de 2022 a dezembro de 2024. Inicialmente, observa-se um aumento gradual dos registros até meados de 2022, seguido de uma estabilização relativa, com médias mensais aproximadas de 500 mil registros. Contudo, notam-se picos significativos em momentos específicos, como em fevereiro de 2024, quando os registros ultrapassam 900 mil, seguido por uma queda acentuada. Esses eventos sugerem a influência de possíveis fatores externos, tais como alterações regulatórias ou variações econômicas, que impactaram de forma significativa o volume de registros.

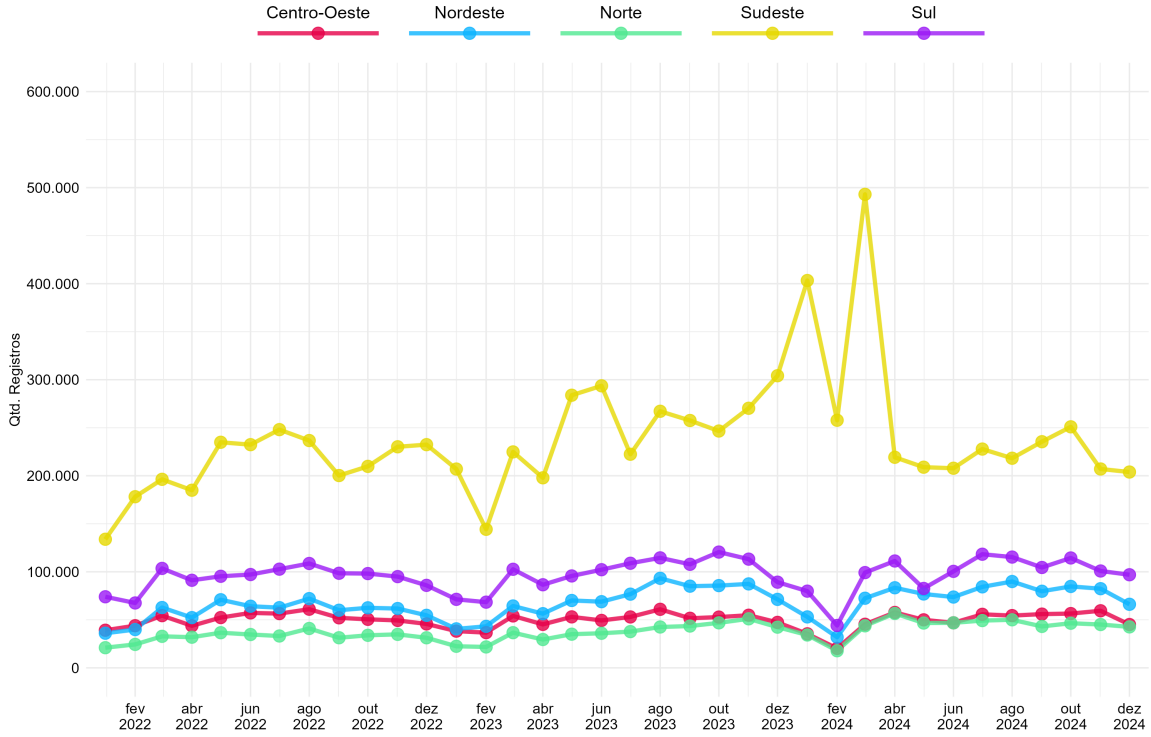
Figura 5.1 Total de registros por mês, 2022 até 2024.



Fonte: Elaboração própria.

A segmentação dos registros por região, apresentada na Figura 5.2, revela disparidades significativas. A região Sudeste destaca-se consistentemente, com volumes que frequentemente excedem 400 mil registros mensais, atingindo mais de 500 mil no auge de fevereiro de 2024. Em contraste, regiões como Norte e Centro-Oeste exibem médias mensais inferiores a 100 mil registros, refletindo a disparidade na concentração econômica. As regiões Sul e Nordeste localizam-se em posições intermediárias, porém com padrões distintos: enquanto o Sul demonstra relativa estabilidade ao longo do período, o Nordeste apresenta maior variabilidade, evidenciando oscilações mensais mais acentuadas.

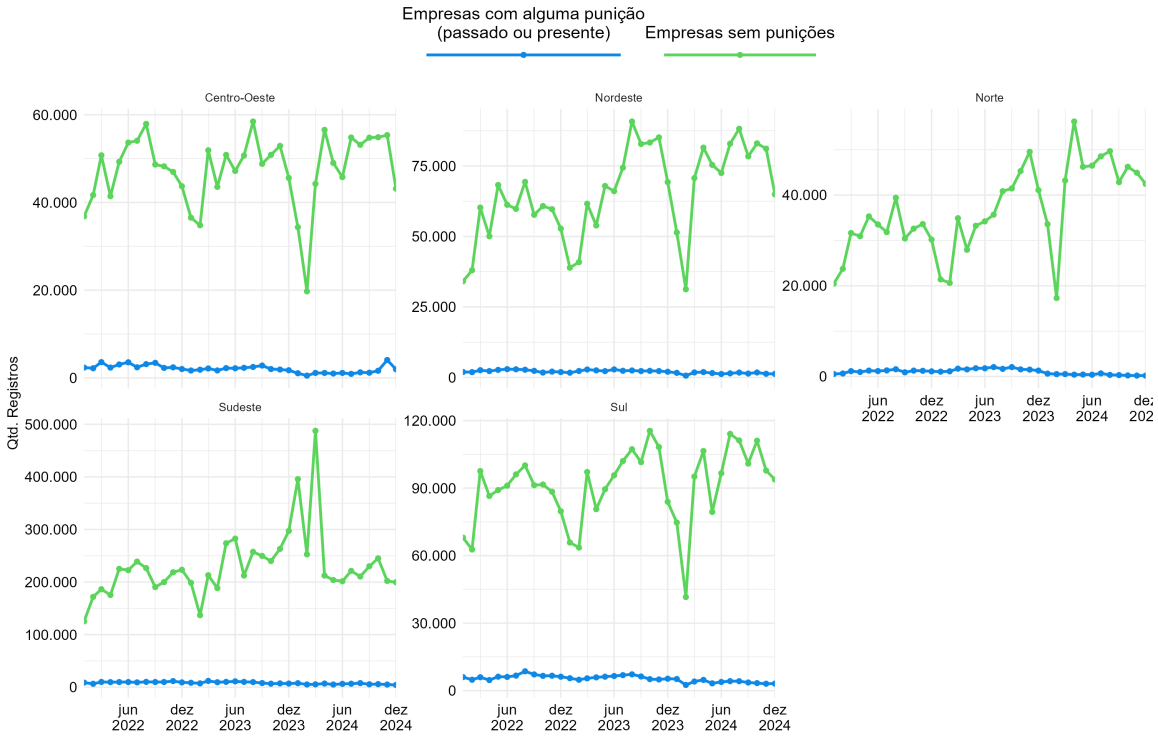
Figura 5.2 Total de registros por região e mês, 2022 até 2024.



Fonte: Elaboração própria.

Ao considerar a classificação das empresas em punidas e não punidas, conforme ilustrado na Figura 5.3, observa-se uma clara predominância das empresas não punidas em todas as regiões analisadas. No Sudeste, por exemplo, os registros mensais de empresas não punidas frequentemente ultrapassam 400 mil, enquanto os registros de empresas punidas raramente superam 10 mil. Tendência semelhante é identificada nas regiões Sul e Nordeste, ainda que em volumes inferiores.

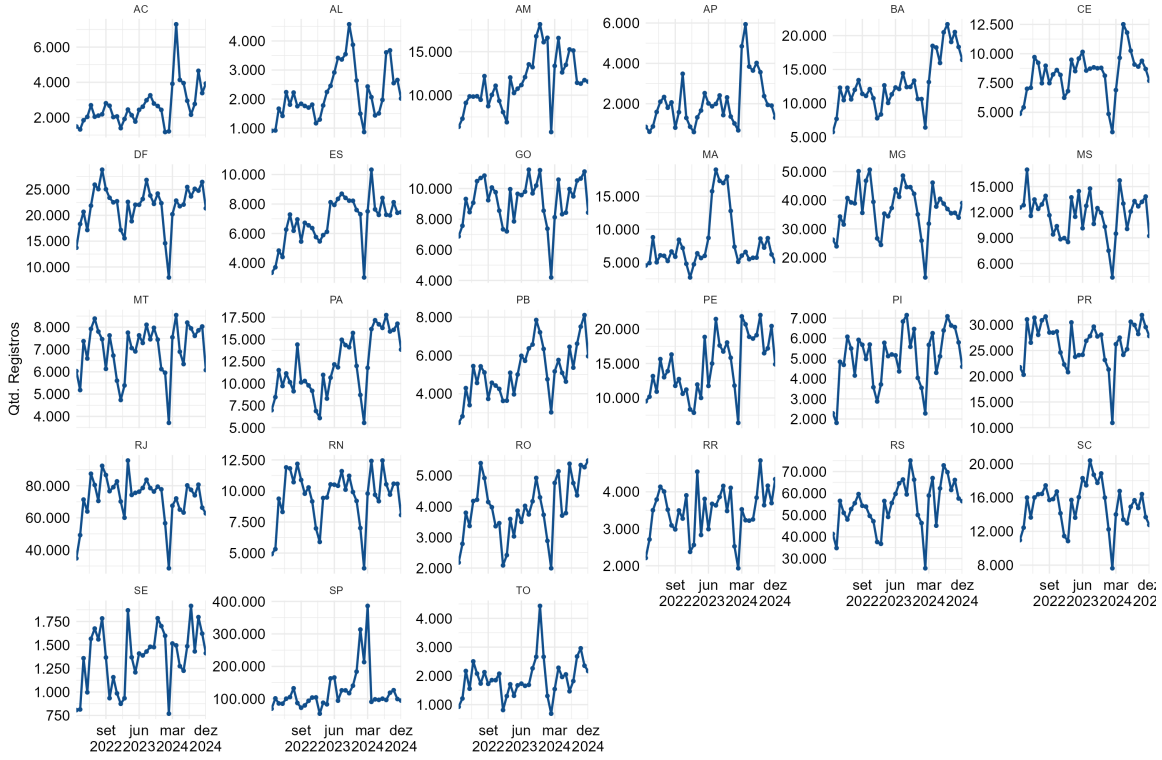
Figura 5.3 Total de registros por região e mês, 2022 até 2024.



Fonte: Elaboração própria.

A análise conduzida por unidade federativa (UF), como ilustrado na Figura 5.4, aprimora a compreensão das variações regionais. São Paulo sobressai como o estado com o maior número absoluto de registros, frequentemente ultrapassando 300 mil por mês. Minas Gerais e Rio de Janeiro apresentam volumes consideráveis, embora significativamente menores quando comparados a São Paulo. Em contrapartida, estados como Acre, Amapá e Roraima exibem os menores volumes, frequentemente apresentando menos de 5 mil registros mensais. A estabilidade apresentada pelos estados de maior porte contrasta com a volatilidade mais acentuada verificada em estados de menor tamanho, como Roraima, onde picos e quedas abruptas indicam a presença de fatores pontuais ou uma menor regularidade na emissão de notas fiscais.

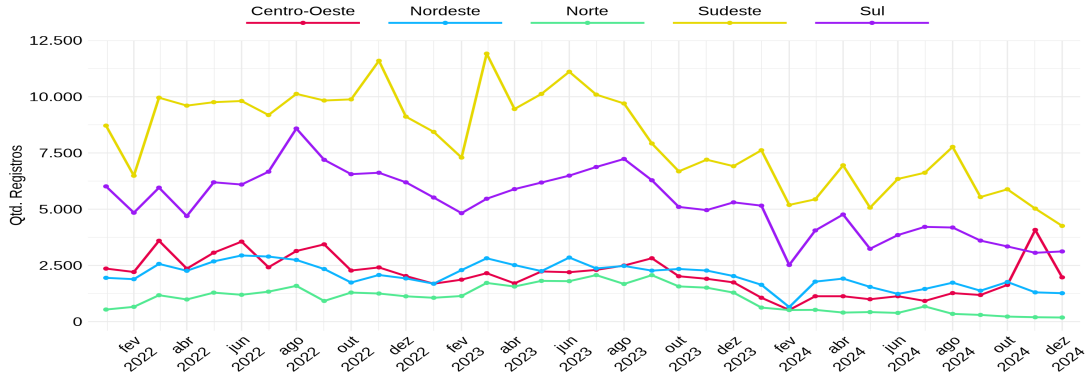
Figura 5.4 Total de registros por unidade federativa (UF) e mês, 2022 até 2024.



Fonte: Elaboração própria.

Os registros exclusivos de empresas penalizadas por região, conforme detalhados na Figura 5.5, reafirmam o protagonismo da região Sudeste, que apresenta o maior volume absoluto de ocorrências em quase todos os períodos analisados. A região Sul ocupa a segunda posição em termos de volumes de registros de empresas penalizadas, enquanto as demais regiões se mantêm em níveis inferiores, salientando-se a região Norte.

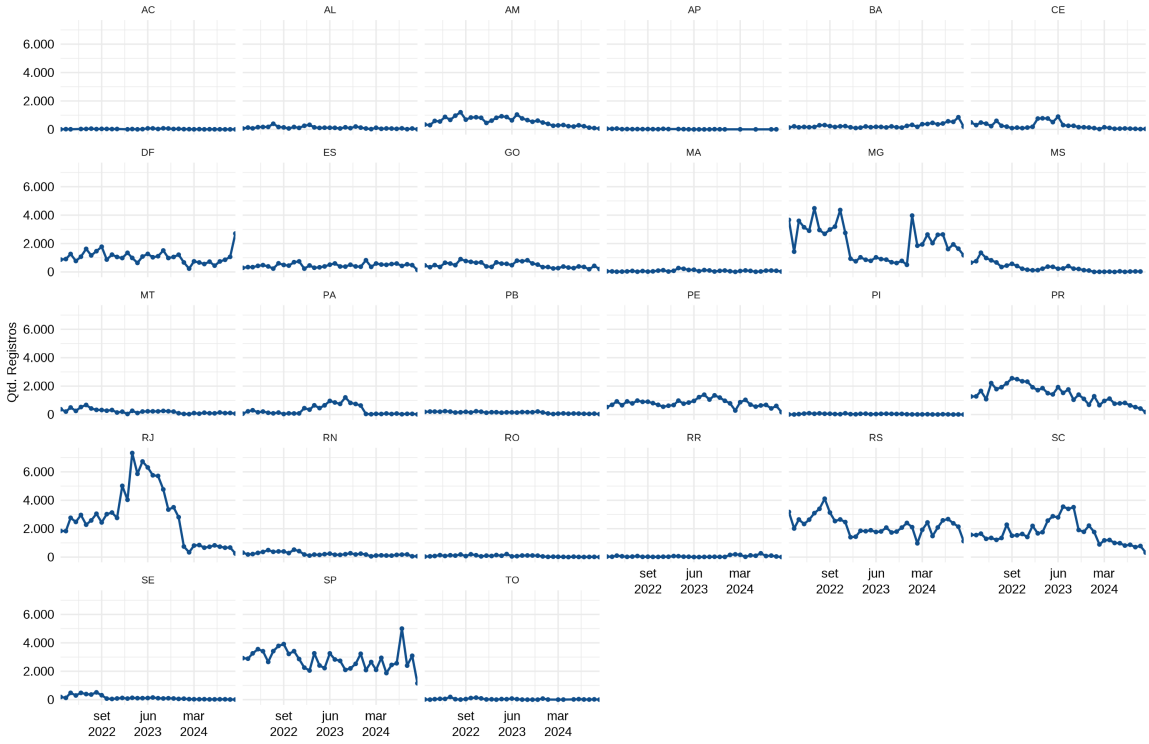
Figura 5.5 Total de registros com empresas que sofreram punições (passadas ou presentes) por região e mês, 2022 até 2024.



Fonte: Elaboração própria.

Finalmente, os dados referentes a empresas penalizadas por unidade federativa, conforme demonstrado na Figura 5.6, evidenciam a preeminência de São Paulo, seguido por estados como Rio de Janeiro e Minas Gerais. Regiões de menor porte, como Acre e Amapá, apresentam volumes significativamente reduzidos e, em diversos períodos, não registram empresas penalizadas. Essa distribuição assimétrica reflete a concentração econômica e a centralização das atividades empresariais em estados mais desenvolvidos, mas também suscita questões sobre a uniformidade na aplicação de sanções fiscais no país. Em estados de menor dimensão, a baixa ocorrência de registros de empresas penalizadas pode indicar tanto uma fiscalização menos intensiva quanto um perfil econômico diferenciado, caracterizado por menor complexidade nas operações empresariais.

Figura 5.6 Total de registros com empresas que sofreram punições (passadas ou presentes) por unidade federativa (UF) e mês, 2022 até 2024.



Fonte: Elaboração própria.

Os resultados apresentados pelos gráficos revelam padrões claros nas emissões de notas fiscais ao longo do tempo e por região, além de destacarem a relevância da classificação das empresas como punidas ou não. Essas análises permitem identificar diferenças estruturais entre regiões e estados, reforçando a importância de compreender fatores específicos que influenciam os volumes de registros e a aplicação de sanções fiscais.

Ao analisar o conjunto de dados empregado no treinamento do modelo, correspondente ao dataset do ano de 2023, a Tabela 5.1 apresenta as estatísticas descritivas das variáveis numéricas referentes à quantidade, ao valor unitário e ao valor total de cada

produto contido no conjunto de dados. As estatísticas apresentadas incluem contagem, média, desvio-padrão, bem como valores mínimos, máximos e quartis. A média para a variável "Quantidade" é de 729,845, ao passo que a média para "Valor Unitário" é de R$ 11.502,250. Estes valores médios sugerem uma grande variabilidade nos dados, especialmente em relação ao "Valor Unitário" e ao "Valor Total", como é demonstrado pelos elevados desvios-padrão de 890.369,232 e 1.014.769,662, respectivamente.

Tabela 5.1 Estatísticas descritivas - 2023

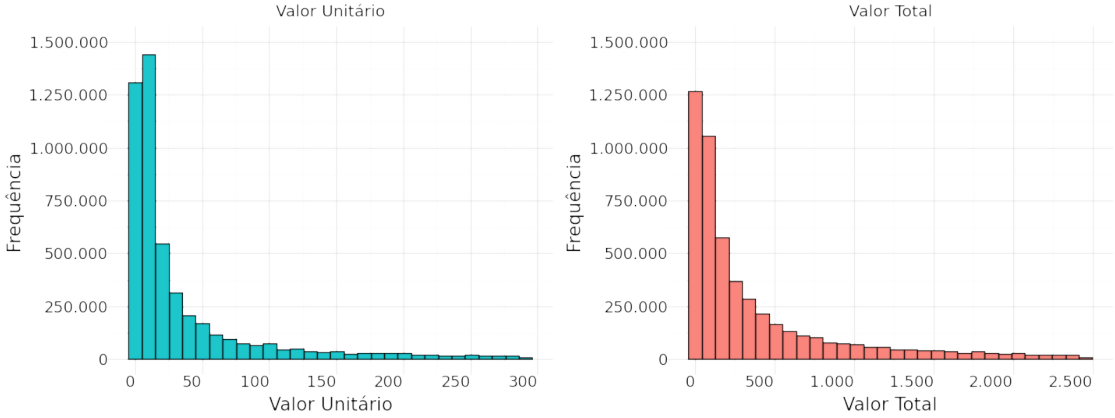
Estatística	**Variáveis**		
	Quantidade	**Valor Unitário**	**Valor Total**
Contagem	5.986.353	5.986.353	5.986.353
Média	729,845	3.598,625	11.502,250
Desvio Padrão	48.826,606	890.369,232	1.014.769,662
Valor mínimo	0	0	0
1 Quartil	1,000	5,890	55,000
2 Quartil	5,000	18,500	230,560
3 Quartil	37,000	120,000	1.015,000
Valor máximo	77.991.000,000	1.257.853.390,600	1.257.853.390,600

Fonte: Elaboração própria.

Os valores mínimos para todas as colunas são 0, o que indica a presença de entradas não significativas ou possivelmente um erro no preenchimento do valor. Os quartis revelam mais sobre a distribuição dos dados: o primeiro quartil para "Quantidade" é 1,000, para "Valor Unitário" é 5,890 e para "Valor Total" é 55,000. Esses valores aumentam significativamente nos quartis seguintes, com o terceiro quartil atingindo 37,000 para "Quantidade", 100,000 para "Valor Unitário" e 1.015,000 para "Valor Total". Os valores máximos são extremamente altos, com "Quantidade" atingindo 77.991.000, "Valor Unitário" chegando a 1.257.853.390,600 e "Valor Total" também alcançando 1.257.853.390,600, indicando a presença de outliers significativos.

A Figura 5.7 ilustra a distribuição dos valores das variáveis de preço, o valor unitário e o valor total, por meio de histogramas. Ambos apresentam uma distribuição assimétrica positiva, com alta concentração de valores próximos de 0 e uma cauda longa que se estende para valores maiores. No histograma de valor unitário, percebe-se um pico de frequência em valores abaixo de 50, com ocorrência muito alta em valores pequenos e redução gradual conforme esses valores aumentam. De maneira similar, o histograma de valor total também evidencia concentração elevada em valores baixos, diminuindo de forma constante na frequência à medida que os valores crescem. Essas distribuições indicam que muitos produtos possuem valores unitários e totais baixos, mas há alguns com valores extremamente elevados que influenciam a média e o desvio padrão.

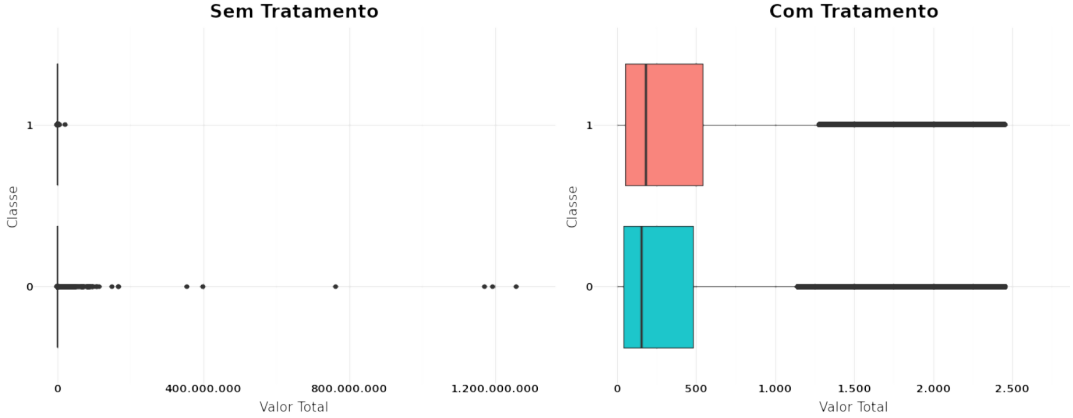
Figura 5.7 Histograma de preço total e unitário - 2023



Fonte: Elaboração própria.

A análise de outliers, apresentada na Figura 5.8, mostra gráficos de boxplots para a variável referente ao valor total e segmentada de acordo com o status da empresa responsável pelo produto, onde 0 se refere a empresas que nunca foram classificadas como inidôneas pelo Governo Federal e 1 para empresas que já receberam (ou ainda mantém) o status de inidônea. Foram representados dois tipos de boxplots, com e sem tratamento. O gráfico "Com Tratamento" passou por um tratamento de remoção parcial de outliers utilizando o método IQR antes de gerar o boxplot. No gráfico sem o tratamento há uma clara presença de valores muito extremos (valores incluindo 1.2 bilhões) que distorcem a representação gráfica dos dados. Quando esses valores muito extremos são removidos, o gráfico de boxplot revela uma distribuição mais compacta e uniforme dos dados, facilitando a interpretação das diferenças entre os quartis. Nota-se que não existe uma diferença elevada entre as classes.

Figura 5.8 Boxplot do "valor total" com e sem tratamento - 2023



Fonte: Elaboração própria.

A nuvem de palavras apresentada na Figura 5.9, gerada a partir da descrição dos produtos, destaca os termos mais frequentes no conjunto de dados. Palavras como “PACOM”, “NOVO”, “BRASAO” e “DPE” aparecem em destaque, indicando sua alta frequência. A presença recorrente desses termos sugere possíveis padrões de descrição ou categorias de produtos predominantes. Além disso, a diversidade de termos menores na nuvem de palavras sugere uma variedade de produtos diferentes descritos no conjunto de dados, refletindo a heterogeneidade dos produtos analisados.

Figura 5.9 Nuvem de palavras da descrição dos produtos

Fonte: Elaboração própria.

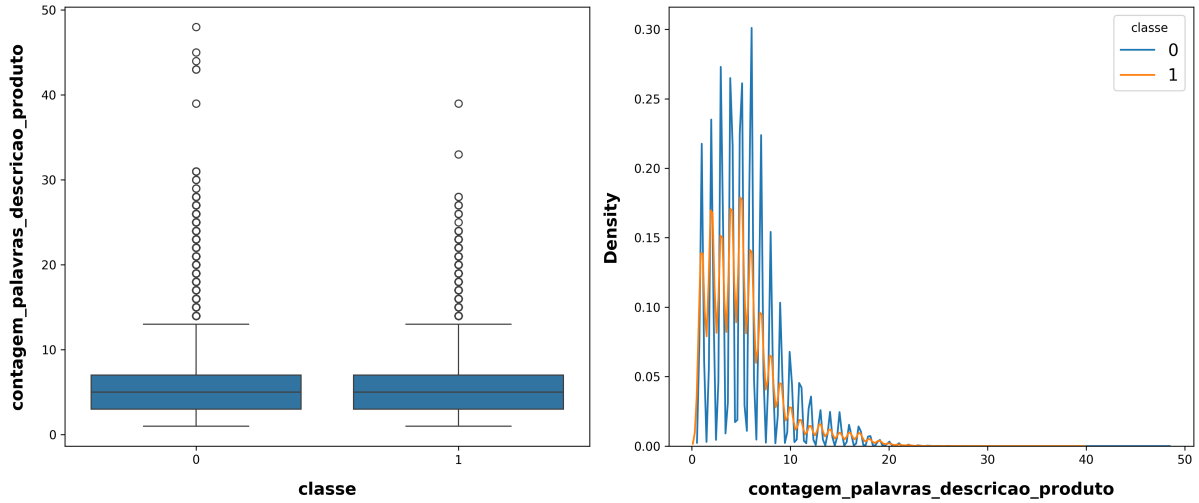
Em conclusão, a análise descritiva dos dados revelam uma grande variabilidade nos valores unitários e totais dos produtos, com a presença de outliers significativos que influenciam as medidas de tendência central. As distribuições assimétricas dos dados e a alta frequência de termos específicos nas descrições dos produtos fornecem insights importantes sobre a natureza dos dados.

A segunda etapa da análise concentra-se na descrição dos produtos, as Figuras 5.10 e 5.11 oferecem uma análise das distribuições da contagem de palavras na descrição do produto e do tamanho da descrição, categorizadas pela variável classe, que é rotulada como 0 e 1.

Na Figura 5.10, à esquerda, o boxplot mostra a distribuição da contagem de palavras na descrição do produto para as duas classes. Observa-se que ambas as classes apresentam uma distribuição similar, com a maioria das descrições contendo entre 5 e 15 palavras. A mediana para ambas as classes é de aproximadamente 10 palavras. No entanto, existem vários outliers acima de 30 palavras, e alguns extremos chegam a quase 50 palavras. A presença desses outliers sugere que, embora a maioria das descrições seja relativamente curta, há exceções significativas de produtos com descrições bem detalhistas.

À direita da Figura 5.10, a densidade da contagem de palavras na descrição do produto é apresentada. As curvas de densidade indicam que a distribuição é levemente assimétrica, com um pico em torno de 5 a 10 palavras e uma cauda longa e mais achatada se estendendo para descrições mais longas. As distribuições das duas classes são bastante semelhantes, embora haja pequenas variações na densidade, sugerindo que a contagem de palavras nas descrições não varia significativamente entre as classes.

Figura 5.10 Boxplot e Densidade da contagem de palavras na descrição dos produtos, por classe - 2023



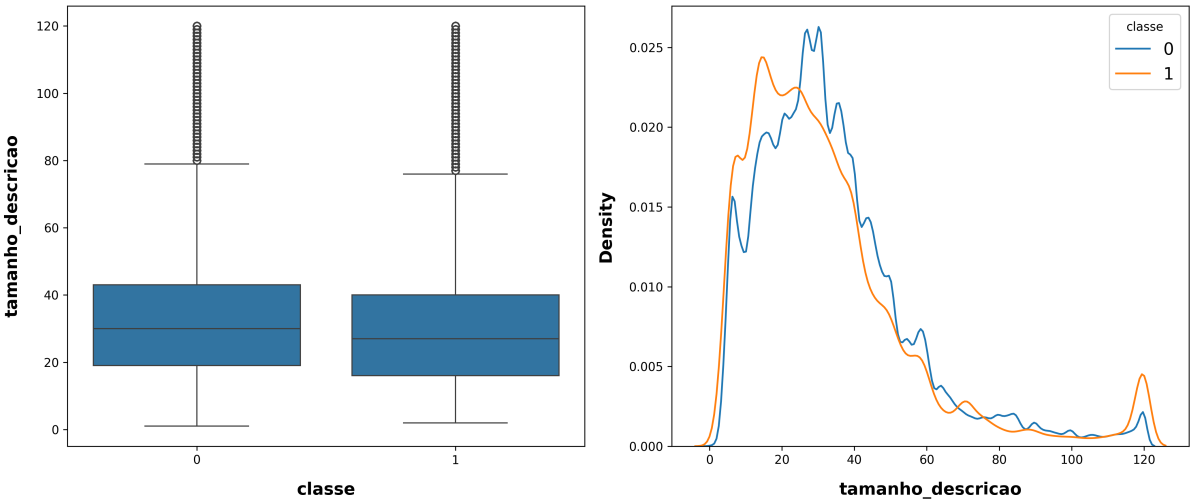
Fonte: Elaboração própria.

Na Figura 5.11, à esquerda, o boxplot apresenta a distribuição do tamanho da descrição em caracteres para as duas classes. Aqui, também se observa uma similaridade entre as classes, com a maioria das descrições variando entre 20 e 60 caracteres. A mediana do tamanho da descrição é cerca de 40 caracteres para ambas as classes. Notam-se vários outliers acima de 80 caracteres, com alguns extremos atingindo até 120 caracteres. O que se observa é que embora a maioria das descrições tenha um tamanho moderado, existem descrições ou palavras significativamente mais longas.

À direita da Figura 5.11 observa-se a densidade do tamanho da descrição em caracteres. As curvas de densidade mostram que a distribuição é bimodal, com picos em torno de 20 a 40 caracteres e outro pico, esse menor, em torno de 100 a 120 caracteres. Isso sugere que, além da maioria das descrições serem curtas a moderadas em tamanho, há um grupo distinto de descrições muito longas. As distribuições para as duas classes são bastante similares, embora existam pequenas diferenças na densidade em certos pontos, indicando uma leve variação no tamanho das descrições entre as classes.

Essas análises detalham como as descrições dos produtos variam em termos de contagem de palavras e tamanho. A similaridade nas distribuições entre as duas classes sugere que a contagem de palavras e o tamanho das descrições não são fortemente diferenciadores entre as classes, embora seja notável a presença de outliers, que podem influenciar na classificação do modelo, esses valores estão presentes em ambas às classes.

Figura 5.11 Boxplot e Densidade do tamanho de cada descrição de produto, por classe - 2023



Fonte: Elaboração própria.

5.2 MODELAGEM DOS DADOS

5.2.1 Treinamento

O procedimento de treinamento inicia com a elaboração de um subconjunto balanceado, composto por N instâncias selecionadas, sendo metade pertencente à classe 0 e a outra metade à classe 1. Posteriormente, este conjunto é estratificadamente particionado para contemplar as três estratégias avaliadas: *zero-shot*, *fine-tuning* e *few-shot*. Cada método é, então, submetido ao treinamento e avaliação, com o registro das métricas de acurácia, precisão, *recall*, F1-score, AUC, matriz de confusão e custo. No cenário de *fine-tuning*, os experimentos são realizados de forma repetida com 3 e 5 épocas; enquanto que para o cenário *few-shot*, o modelo recebe amostras de 10 e 50 exemplos de cada classe.

5.2.1.1 *Zero-Shot*

De acordo com Wang, Pang e Lin (2023), *Zero-shot learning* (ZSL), trata-se de uma abordagem inovadora que possibilita a classificação de dados sem a exigência de exemplos específicos de treinamento para determinadas classes. Tal técnica adota modelos pré-treinados capazes de prever tanto classes conhecidas quanto desconhecidas, fundamentando-se em instruções textuais. A característica fundamental do ZSL reside em sua capacidade de operar sem etapas intermediárias, como a extração de características ou tokenização, o que simplifica o processo e reduz os custos computacionais (WANG; PANG; LIN, 2023). No contexto de aprendizado de máquina e processamento de linguagem natural (PLN), o ZSL elimina a dependência de dados amplamente rotulados, permitindo que as tarefas sejam executadas a partir de descrições, sem necessidade de *fine-tuning* ou reconfiguração do modelo apropriado.

Figura 5.12 Comparativo entre pipeline tradicional de treinamentos e treinamento *zero-shot*

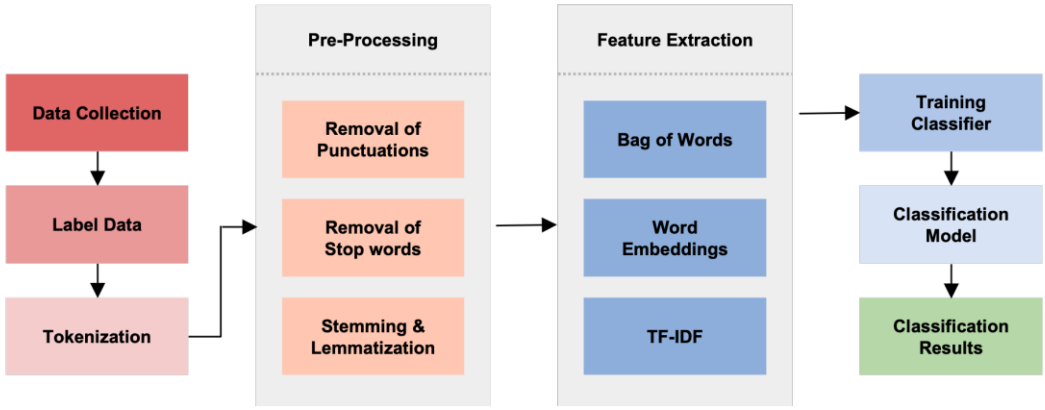


Fig. 1. Traditional text classification flow

Fig. 2. LLMs' zero shot text classification flow

Fonte: Extraído de Wang, Pang e Lin (2023).

Ao explorar a aplicação do *Zero-Shot Learning* (ZSL) no campo do processamento de linguagem natural, conforme mencionado por Wang, Pang e Lin (2023), o ZSL revelou-se crucial ao possibilitar a classificação de textos sem a necessidade de treinamento específico. Modelos de linguagem de grande porte (LLMs), como GPT-3.5, GPT-4 e Llama2, exemplificam essa implementação. Tais modelos empregam técnicas de "prompting" para orientar o processamento e produzir resultados precisos em diversas tarefas, tais como análise de sentimentos, detecção de spam e classificação de textos. Além disso, conforme citado em Sivarajkumar e Wang (2023), o ZSL apresenta um potencial significativo, especialmente em contextos desafiadores, como o domínio clínico, onde a extração de entidades nomeadas e a classificação de textos enfrentam severas limitações devido à escassez de dados rotulados disponíveis.

Tabela 5.2 Comparação de modelos em configuração *zero-shot*

Modelo	Amostra			Métricas de Desempenho					Matriz de Confusão				Matriz de Custo	
	T. Total	Label 1	Label 0	Acurácia	Precisão	*Recall*	F1	AUC	TN	FP	FN	TP	FP=1 / FN=5	FP=5 / FN=1
openai-community/gpt2-large	20.000	10.000	10.000	0,4925	0,4962	0,9634	0,6550	0,5028	217	9.783	366	9.634	0,5806	2,4640
mistralai/Mistral-7B-Instruct-v0.3	20.000	10.000	10.000	0,4967	0,4783	0,0729	0,1265	0,5000	9.205	795	9.271	729	2,3575	0,6623
google/gemma-2b-it	20.000	10.000	10.000	0,4928	0,4963	0,9585	0,6540	0,4815	271	9.729	415	9.585	0,5902	2,4530
meta-llama/Llama-3.1-8B	20.000	10.000	10.000	0,4859	0,4916	0,8232	0,6156	0,4677	1.486	8.514	1.768	8.232	0,8677	2,2169
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	20.000	10.000	10.000	0,4938	0,4619	0,7570	0,1301	0,4433	9.118	882	9243	757	2,3548	0,6826

Fonte: Elaboração própria.

Na análise da Tabela 5.2, observa-se uma variação expressiva no desempenho dos modelos avaliados em um cenário *zero-shot*, particularmente na relação entre falsos positivos (FP) e falsos negativos (FN), refletindo diretamente sobre o custo ponderado das

classificações. Modelos como o "openai-community/gpt2-large" e "google/gemma-2b-it" destacam-se por atingir recalls muito elevados (acima de 95%), com F1-Scores em torno de 0,65. Contudo, o preço dessa sensibilidade é uma quantidade excessiva de falsos positivos, levando a custos altos quando os FPs têm maior peso na análise.

Em contrapartida, modelos como "mistralai/Mistral-7B-Instruct-v0.3" e "deepseek-ai/DeepSeek-RL-Distill-Qwen-7B" adotam uma postura oposta, reduzindo drasticamente os falsos positivos, mas comprometendo o recall a níveis críticos (abaixo de 8%), resultando em valores extremamente baixos de F1-Score (cerca de 0,13). O modelo "meta-llama/Llama-3-8B" se posiciona em um meio-termo, alcançando equilíbrio relativamente melhor entre as duas métricas, embora ainda não ideal. Isso evidencia que, no cenário *zero-shot*, nenhum dos modelos avaliados consegue simultaneamente manter níveis satisfatórios de precisão e recall.

O prompt apresentado na Figura 5.13 foi estruturado com instruções claras e objetivas para a tarefa de classificação, fornecendo contexto explícito ao modelo sobre sua função como auditor especializado. A definição precisa das classes (0 para compras normais e 1 para compras suspeitas) visou reduzir ambiguidades e facilitar a interpretação correta pelo modelo. No entanto, mesmo com essa formulação, os resultados demonstram que a técnica *zero-shot* é limitada pela tendência geral dos modelos em produzir classificações polarizadas, destacando a necessidade de abordagens mais específicas para alcançar um desempenho prático satisfatório.

Figura 5.13 Prompt para classificação *zero-shot*

[QUEM É VOCÊ]
Você é um Auditor Especializado em detectar empresas potencialmente não confiáveis por meio de extratos de compras.

[O QUE VOCÊ FAZ]
Classifica, com base na descrição do extrato de compra, em duas categorias:

- **0 (Normal)** — compra rotineira e claramente descrita;
- **1 (Suspeita)** — descrição vaga, item de luxo ou produto/serviço inadequado para a entidade pública.

[TEXTO PARA CLASSIFICAR]

```
Texto: {text}
```

[CLASSIFICAÇÃO]
Classificação:

5.2.1.2 *Fine-Tuning*

O processo de *fine-tuning* envolve utilizar as configurações de um modelo pré-treinado e ajustá-los para uma tarefa específica, como a classificação de textos. Durante o *fine-tuning*, a camada superior do modelo é modificada para se adequar ao espaço de rótulos e às perdas da tarefa específica, após isso, tanto os novos parâmetros quanto os originais são co-treinados (HOULSBY et al., 2019). Ainda segundo Houlsby et al. (2019), o *fine-tuning* tem mostrado resultados de estado da arte em várias tarefas de PLN, como classificação de textos e pergunta/resposta. O *fine-tuning* frequentemente apresenta melhor desempenho do que a transferência baseada em características (HOWARD; RUDER, 2018), um método baseado no uso de recursos extraídos na tarefa de origem como ponto de partida para a tarefa de destino.

A taxa de aprendizado é um hiper-parâmetro que controla o tamanho dos passos das atualizações dos parâmetros do modelo durante o treinamento. Uma taxa de aprendizado menor permite ajustes mais precisos nos pesos do modelo, o que é particularmente importante no *fine-tuning* de modelos pré-treinados, para evitar a perda do conhecimento pré-treinado (MOSBACH; ANDRIUSHCHENKO; KLAKOW, 2021). A taxa de aprendizado utilizada no treinamento foi de 2e-5. Esse valor foi selecionado com base em resultados empíricos apresentados por Mosbach, Andriushchenko e Klakow (2021), onde os autores evidenciam que esse valor proporciona um bom equilíbrio entre a velocidade de convergência e a estabilidade do treinamento.

O valor do decaimento de peso, que é uma técnica de regularização usada para prevenir overfitting, penalizando pesos grandes no modelo, utilizado foi de 0.1. Segundo Mosbach, Andriushchenko e Klakow (2021), esse valor equilibra a necessidade de regularização com a capacidade do modelo de se adaptar à tarefas específicas durante o *fine-tuning*. A regularização ajuda a manter a capacidade de generalização do modelo, evitando que ele se ajuste muito aos dados de treinamento e, assim, melhorando seu desempenho em dados não vistos.

O pipeline de treinamento de modelo é apresentando na Figura 5.14; O processo inicia-se com o dataset original. Na primeira etapa, esse conjunto é balanceado para manter 50% de exemplos por classe, minimizando vieses de prevalência e tornando as métricas comparáveis entre grupos.

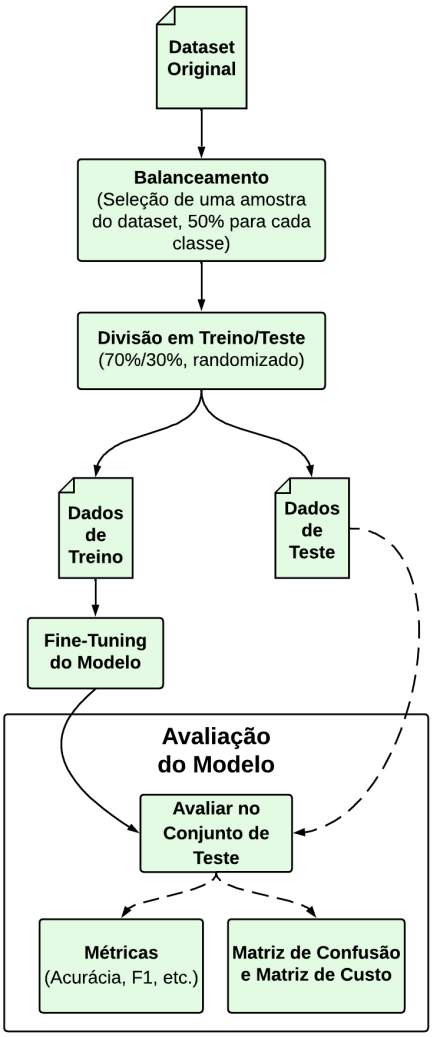
Em seguida, o fluxo executa uma separação aleatória 70%/30% em dados de treino e de teste, mantendo um balanceamento entre as duas classes. Essa divisão garante que o ajuste do modelo ocorra apenas sobre 70% dos exemplos, preservando os 30% restantes como referência imparcial para mensurar desempenho.

Os dados de treino alimentam o *fine-tuning*, estágio em que hiper-parâmetros e pesos do algoritmo são ajustados. Concluído o ajuste, o modelo é passado à fase de avaliação, na qual é testado exclusivamente no conjunto de teste. Daqui resultam métricas globais (acurácia, F1, entre outras), a matriz de confusão e a matriz de custo, que detalham a natureza e o impacto dos erros de classificação.

A Tabela 5.3 apresenta os resultados do treinamento utilizando a técnica de *fine-tuning* para cada modelo avaliado. Em termos gerais, no grupo ajustado por apenas três épocas, o "google/gemma-2-2b" desponta como o ponto de referência: combina acurácia

e F1 superiores às versões equivalentes de Llama e apresenta o menor custo médio de erro ($\approx$ 0,28 para FN = 5 e 0,31 para FP = 5), sinalizando bom equilíbrio entre falsos positivos e negativos. Os dois modelos meta-llama com 7b e 3,2B, embora muito próximos, sofrem levemente com um aumento da taxa de falsos positivos, diferença que se reflete em custos algo maiores. Em contraste, os sistemas baseados em GPT-2 mostram as maiores fragilidades desse primeiro bloco: perdem precisão de modo sensível, o que empurra o custo ponderado a patamares três a quatro vezes mais altos do que o observado no gemma.

Figura 5.14 Treinamento do modelo com *fine-tuning*



Fonte: Elaboração própria.

Quando se amplia o treinamento para cinco épocas, com o respectivo acréscimo de amostras, o panorama muda de forma significativa. O próprio gemma-2-2b eleva todos os indicadores e passa a operar com custo inferior a 0,20 para FN e 0,23 para FP, mas deixa de ser o extremo superior da distribuição. A versão de 40.000 exemplos do meta-llama 3,2 B ultrapassa o gemma em acurácia e produz a melhor área sob a curva entre os Llama,

entretanto é superada por dois modelos recém-introduzidos: os da família DeepSeek. O "DeepSeek-R1-Distill-Qwen-7B" apresenta F1 superior a 0,93 e custo perto de 0,19 para FN e 0,20 para FP, mas o maior destaque recai sobre o "DeepSeek-R1-Distill-Llama-8B", que atinge F1 $\approx$ 0,95 e reduz o custo ponderado para a faixa de 0,13–0,16, tornando-se o modelo com o menor impacto de erro da tabela. Vale notar que, para esses modelos, tanto precisão quanto *recall* permanecem acima de 0,94, sinalizando que os ganhos não são obtidos à custa de sensibilidade ou especificidade.

De maneira geral, verifica-se que o incremento no número de épocas e na quantidade de dados contribui significativamente para a estabilidade do modelo, melhorando as métricas de desempenho e mitigando os impactos adversos dos erros de classificação, com os modelos "deepseek-ai/DeepSeek-R1-Distill-Llama-8B" e "deepseek-ai/DeepSeek-R1-Distill-Qwen-7B" e "meta-llama/Llama-2-7b-hf" destacando-se entre os demais nos cenários analisados.

Tabela 5.3 Comparação de modelos em configuração *Fine-Tuning*

Modelo	Epochs	Amostra			Métricas de Desempenho					Matriz de Confusão				Matriz de Custo	
		T. Total	Label 1	Label 0	Acurácia	Precisão	*Recall*	F1	AUC	TN	FP	FN	TP	FP=1 / FN=5	FP=5 / FN=1
meta-llama/Llama-2-7b-hf	3	20.000	10.000	10.000	0,8990	0,8970	0,9004	0,8988	0,9554	2.700	308	298	2.694	0,29967	0,30633
meta-llama/Llama-3.2-3B	3	20.000	10.000	10.000	0,8841	0,8742	0,8967	0,8853	0,9400	2.622	386	309	2.683	0,32183	0,37317
google/gemma-2-2b	3	20.000	10.000	10.000	0,9020	0,8948	0,9104	0,9025	0,9543	2.688	320	268	2.724	0,27667	0,31133
openai-community/gpt2	3	20.000	10.000	10.000	0,6801	0,6286	0,8763	0,7320	0,7612	1.459	1.549	370	2.622	0,56650	1,35250
openai-community/gpt2-large	3	20.000	10.000	10.000	0,7700	0,7185	0,8856	0,7934	0,8729	1.970	1.038	342	2.650	0,45800	0,92200
neuralmind/bert-base-portuguese-cased	3	20.000	10.000	10.000	0,7548	0,7023	0,8823	0,7821	0,8492	1.889	1.119	352	2.640	0,47983	0,99117
neuralmind/bert-large-portuguese-cased	3	20.000	10.000	10.000	0,7738	0,7192	0,8963	0,7980	0,8592	1.961	1.047	310	2.682	0,43283	0,92417
google/gemma-2-2b	5	40.000	20.000	20.000	0,9279	0,9205	0,9383	0,9293	0,9683	5.446	491	374	5.689	0,19675	0,23575
neuralmind/bert-large-portuguese-cased	5	40.000	20.000	20.000	0,8630	0,8218	0,9307	0,8729	0,9483	4.714	1.223	420	5.643	0,27692	0,54458
meta-llama/Llama-3.2-3B	5	40.000	20.000	20.000	0,9398	0,9345	0,9472	0,9408	0,9786	5.535	402	320	5.743	0,16683	0,19417
meta-llama/Llama-2-7b-hf	5	40.000	20.000	20.000	0,9487	0,9409	0,9587	0,9497	0,9810	5.572	365	250	5.813	0,13458	0,17292
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	5	40.000	20.000	20.000	0,9336	0,9323	0,9366	0,9345	0,9745	5.525	412	384	5.679	0,19433	0,20367
deepseek-ai/DeepSeek-R1-Distill-Llama-8B	5	40.000	20.000	20.000	0,9519	0,9457	0,9599	0,9527	0,9812	5.603	334	243	5.820	0,12908	0,15942

Fonte: Elaboração própria.

5.2.1.3 *Few-Shot*

A técnica *few-shot* representa um método de aprendizado que capacita o modelo a executar tarefas específicas mediante um número reduzido de exemplos fornecidos no contexto da interação, sem necessidade de atualizações nos pesos do modelo. Este método contrasta com abordagens tradicionais que requerem grandes conjuntos de dados e técnicas de ajustes específicos, como o *fine-tuning* (BROWN et al., 2020). De acordo com Parnami e Lee (2022), na aplicação à classificação de textos, o aprendizado *few-shot* é particularmente relevante, pois diminui a dependência de grandes conjuntos de dados rotulados, especialmente em contextos onde a coleta de dados é onerosa ou sensível, como em questões de privacidade. Nesse sentido, Brown et al. (2020) investigam a capacidade de aprendizado em cenários de *few-shot* em modelos de linguagem, com destaque para o GPT-3, um modelo autorregressivo composto por 175 bilhões de parâmetros. Os experimentos realizados avaliaram o GPT-3 em mais de 25 conjuntos de dados de PLN, abrangendo tarefas como tradução, compreensão de leitura, raciocínio lógico e classificação de texto. No contexto da classificação de textos, o GPT-3 apresentou resultados promissores, demonstrando adaptação a tarefas como inferência lógica e escolha de sentenças corretas com desempenho competitivo, mesmo em cenários *zero-shot* ou *one-shot*.

Nos testes realizados com configuração *few-shot* (Tabela 5.4), observou-se que o desempenho dos modelos foi sensível tanto à quantidade de exemplos fornecidos quanto ao viés intrínseco de cada arquitetura em relação às classes. Alguns modelos apresentaram comportamentos extremos: o "google/gemma-3-12b-pt" e o "deepseek-ai/DeepSeek-R1-Distill-Qwen-7B", por exemplo, com apenas 10 exemplos por classe, tiveram desempenho próximo à aleatoriedade, classificando quase todas as instâncias como negativas. Este cenário é particularmente evidente no Gemma, que não conseguiu identificar nenhum verdadeiro positivo, deixando métricas como precisão, F1-score e AUC indefinidas, além de elevar significativamente o custo de classificação devido à penalização severa dos falsos negativos.

Tabela 5.4 Comparação de modelos em configuração *few-shot*

Modelo	Qtd. de exemplospara cada classe	Amostra			Métricas de Desempenho					Matriz de Confusão				Matriz de Custo	
		T. Total	Label 1	Label 0	Acurácia	Precisão	*Recall*	F1	AUC	TN	FP	FN	TP	FP=1 / FN=5	FP=5 / FN=1
mistralai/Mistral-7B-v0.1	10/10	20.000	10.000	10.000	0,5158	0,6891	0,0583	0,1075	–	9.732	263	9.422	583	2,36865	0,53685
google/gemma-3-12b-pt	10/10	20.000	10.000	10.000	0,4998	–	–	–	–	9.995	–	10.005	–	2,50125	0,50025
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	10/10	20.000	10.000	10.000	0,5009	0,5809	0,0079	0,1560	–	9.938	57	9.926	79	2,48435	0,51055
meta-llama/Llama-3.2-3B	10/10	20.000	10.000	10.000	0,5031	0,5017	0,9738	0,6622	–	319	9.676	262	9.743	0,54930	2,43210
meta-llama/Llama-3.1-8B	10/10	20.000	10.000	10.000	0,6088	0,6261	0,5412	0,5806	–	6.761	3.234	4.590	5.415	1,30920	1,03800
meta-llama/Llama-3.1-8B	50/50	20.000	10.000	10.000	0,5592	0,6864	0,2220	0,3355	–	8.958	1.017	7.799	2.226	2,00060	0,64420
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	50/50	20.000	10.000	10.000	0,5004	0,6518	0,0073	0,0144	–	9.936	39	9.952	73	2,48995	0,50735

Fonte: Elaboração própria.

Por outro lado, o modelo "meta-llama/Llama-3.2-3B" apresentou comportamento inverso na mesma configuração, com recall elevado (0,97), mas à custa de inúmeros falsos positivos, somente cerca de 300 verdadeiros negativos restaram dentre 20.000 previsões. Esse resultado gerou uma drástica variação no custo conforme a penalidade atribuída aos erros: o custo diminuiu bastante ao penalizar fortemente falsos negativos, mas aumentou consideravelmente ao penalizar falsos positivos.

Dentre esses extremos, destacou-se o modelo "Llama-3.1-8B", também com 10 exemplos por classe, conseguindo equilibrar melhor as classificações. Este modelo alcançou acurácia de 0,61 e F1-score de 0,58, mostrando maior robustez relativa em cenários com poucos exemplos, ao aumentar a quantidade de exemplos para 50 por classe não trouxe melhorias, uma vez que, apesar do aumento de precisão (0,69), o recall diminuiu drasticamente (0,22), derrubando o F1-score para 0,34 e aumentando novamente o custo quando falsos negativos têm penalidade alta. O "DeepSeek-Qwen-7B" também não apresentou ganho com aumento dos exemplos, permanecendo com fraco desempenho na detecção da classe positiva.

O *prompt* utilizado para guiar esses testes (Figura 5.15) estabeleceu o contexto de uma auditoria especializada na identificação de empresas potencialmente não confiáveis com base em descrições de compras. O modelo foi orientado a classificar as compras em duas categorias: 0, quando descrições são rotineiras e detalhadas, ou 1, quando as descrições são vagas, referentes a itens de luxo ou inadequadas para uma entidade pública. O *prompt* é muito similar ao utilizado no *zero-shot*, com um adicional para passar os exemplos que o modelo vai se orientar.

Com base nos resultados obtidos, evidenciou-se que, em contextos *few-shot*, a seleção do modelo e a determinação adequada do número de exemplos são cruciais para assegurar um desempenho satisfatório. A análise indica que, embora a ampliação do número de

exemplos pareça intuitivamente benéfica, tal estratégia não necessariamente resulta em melhorias diretas, sendo igualmente essencial considerar o equilíbrio entre as classes e o viés inerente dos modelos. Por último, constatou-se que nenhum modelo avaliado por meio da técnica *few-shot* produziu resultados satisfatórios para aplicação na tarefa de classificação proposta neste trabalho.

Figura 5.15 Prompt para classificação *few-shot*

[QUEM É VOCÊ]
Você é um Auditor Especializado em detectar empresas potencialmente não confiáveis por meio de extratos de compras.

[O QUE VOCÊ FAZ]
Classifica, com base na descrição do extrato de compra, em duas categorias:

- **0 (Normal)** — compra rotineira e claramente descrita;
- **1 (Suspeita)** — descrição vaga, item de luxo ou produto/serviço inadequado para a entidade pública.

[EXEMPLOS CLASSIFICADOS]
Abaixo estão alguns exemplos:

```
Exemplo: {text}

Classificação: {label}
```

[TEXTO PARA CLASSIFICAR]

```
Texto: {text}
```

[CLASSIFICAÇÃO]
Classificação:

5.3 AVALIAÇÃO DOS RESULTADOS

A Tabela 5.5 sintetiza os resultados obtidos pelos modelos e técnicas avaliadas neste estudo. Primeiramente, é possível observar que as abordagens *zero-shot* não apresentam desempenhos satisfatórios, com acurácia inferior a 0,52 e F1-Scores não ultrapassando 0,66. Embora o modelo "openai-community/gpt2-large" registre o maior *recall* do grupo (0,9634), a baixa precisão (0,4962) resulta em um elevado custo quando falsos positivos são penalizados (2,46405). Por outro lado, modelos como o "mistralai/Mistral-7B-Instruct"

exibem um *recall* extremamente baixo (0,0729), ocasionando altos custos em contextos sensíveis a falsos negativos (custo superior a 2,35).

Na abordagem *few-shot*, o desempenho dos modelos é bastante variável. Com 10 exemplos por classe, o "meta-llama/llama-3.1-8B" obteve o melhor desempenho equilibrado, com acurácia de 0,6088, F1-Score de 0,5806 e *recall* elevado (0,9738), mantendo o custo relativamente baixo (1,03 para FP=5/FN=1). Já o modelo "mistralai/Mistral-7B-v0.1" apresentou alta precisão (0,6891), porém baixo *recall* (0,0583), resultando em F1-Score de apenas 0,1075 e custo elevado quando falsos negativos são caros. Outros modelos, como o "google/gemma-3-12b-pt" e o "deepseek-ai/DeepSeek-R1-Distill-Qwen-7B", praticamente ignoram a classe positiva, refletindo altos custos gerais. Quando o número de exemplos é aumentado para 50 por classe, não se verifica melhora substancial no desempenho. O "meta-llama/llama-3.1-8B" sofre uma queda significativa no *recall* e no F1-Score, enquanto o desempenho do "deepseek-ai/DeepSeek-R1-Distill-Qwen-7B" permanece baixo.

Tabela 5.5 Comparação dos modelos e técnicas avaliadas

Modelo	Técnica	Epochs	Métricas de desempenho					Matriz de Custo - Normalizada	
			Acurácia	Precisão	Recall	F1 Score	AUC	FP = 1 / FN = 5	FP = 5 / FN = 1
openai-community/gpt2-large	Zero-Shot	-	0,4925	0,4962	0,9634	0,6550	0,5028	0,58065	2,46405
mistralai/Mistral-7B-Instruct-v0.3	Zero-Shot	-	0,4967	0,4783	0,0729	0,1265	0,5000	2,35750	0,66230
google/gemma-2b-it	Zero-Shot	-	0,4928	0,4963	0,9585	0,6540	0,4815	0,59020	2,45300
meta-llama/Llama-3.1-8B	Zero-Shot	-	0,4859	0,4916	0,8232	0,6156	0,4677	0,86770	2,21690
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	Zero-Shot	-	0,4938	0,4619	0,7570	0,1301	0,4433	2,35485	0,68265
mistralai/Mistral-7B-v0.1 (10 exemplos)	Few-Shot	-	0,5158	0,6891	0,0583	0,1075	-	2,36865	0,53685
google/gemma-3-12b-pt (10 exemplos)	Few-Shot	-	0,4998	-	-	-	-	2,50125	0,50025
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B (10 exemplos)	Few-Shot	-	0,5009	0,5809	0,0079	0,1560	-	2,48435	0,51055
meta-llama/Llama-3.2-3B (10 exemplos)	Few-Shot	-	0,5031	0,5017	0,9738	0,6622	-	0,54930	2,43210
meta-llama/Llama-3.1-8B (10 exemplos)	Few-Shot	-	0,6088	0,6261	0,5412	0,5806	-	1,30920	1,03800
meta-llama/Llama-3.1-8B (50 exemplos)	Few-Shot	-	0,5592	0,6864	0,2220	0,3355	-	2,00060	0,64420
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B (50 exemplos)	Few-Shot	-	0,5004	0,6518	0,0073	0,0144	-	2,48995	0,50735
meta-llama/Llama-2-7b-hf	Fine-Tuning	3	0,8990	0,8970	0,9004	0,8988	0,9554	0,29967	0,30633
meta-llama/Llama-3.2-3B	Fine-Tuning	3	0,8841	0,8742	0,8967	0,8853	0,9400	0,32183	0,37317
google/gemma-2-2b	Fine-Tuning	3	0,9020	0,8948	0,9104	0,9025	0,9543	0,27667	0,31133
openai-community/gpt2	Fine-Tuning	3	0,6801	0,6286	0,8763	0,7320	0,7612	0,56650	1,35250
openai-community/gpt2-large	Fine-Tuning	3	0,7700	0,7185	0,8856	0,7934	0,8729	0,45800	0,92200
neuralmind/bert-base-portuguese-cased	Fine-Tuning	3	0,7548	0,7023	0,8823	0,7821	0,8492	0,47983	0,99117
neuralmind/bert-large-portuguese-cased	Fine-Tuning	3	0,7738	0,7192	0,8963	0,7980	0,8592	0,43283	0,92417
google/gemma-2-2b	Fine-Tuning	5	0,9279	0,9205	0,9383	0,9293	0,9683	0,19675	0,23575
neuralmind/bert-large-portuguese-cased	Fine-Tuning	5	0,8630	0,8218	0,9307	0,8729	0,9483	0,27692	0,54458
meta-llama/Llama-3.2-3B	Fine-Tuning	5	0,9398	0,9345	0,9472	0,9408	0,9786	0,16683	0,19417
meta-llama/Llama-2-7b-hf	Fine-Tuning	5	0,9487	0,9409	0,9587	0,9497	0,9810	0,13458	0,17292
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	Fine-Tuning	5	0,9336	0,9323	0,9366	0,9345	0,9745	0,19433	0,20367
deepseek-ai/DeepSeek-R1-Distill-Llama-8B	Fine-Tuning	5	0,9519	0,9457	0,9599	0,9527	0,9812	0,12908	0,15942

Fonte: Elaboração própria

Em contrapartida, os métodos de *fine-tuning* apresentam um avanço expressivo em todas as métricas analisadas. Três modelos treinados por cinco épocas destacam-se claramente: o "deepseek-ai/DeepSeek-R1-Distill-Llama-8B", o "meta-llama/llama-2-7b-hf" e o "meta-llama/llama-3.2-3B". Dentre esses, o modelo da DeepSeek alcançou o melhor desempenho geral, com acurácia de 0,9519, F1-Score de 0,9527 e AUC de 0,9812. Além disso, este modelo apresentou o menor custo entre todos os avaliados (0,15942 para FP=5/FN=1). Próximo desse resultado está o "meta-llama/Llama-2-7b-hf", com acurácia de 0,9487, F1-Score de 0,9497 e AUC de 0,9810, mantendo custo igualmente competitivo (0,17297 para FP=5/FN=1). O "meta-llama/llama-3.2-3B", embora um pouco inferior, permanece competitivo, alcançando acurácia de 0,9398 e F1-Score de 0,9408. Como referência adicional, o modelo "google/gemma-2-2b" obteve métricas ligeiramente inferiores após cinco épocas de treinamento (acurácia de 0,9279 e F1-Score de 0,9293),

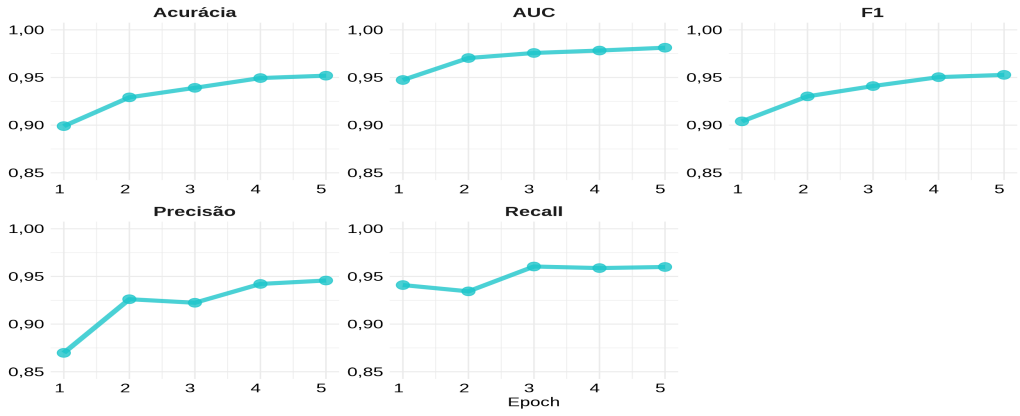
ainda assim significativamente à frente dos modelos baseados em GPT-2 ou BERT, cujo custo é consistentemente superior.

Analisando mais profundamente o modelo "DeepSeek-R1-Distill-Llama-8B" (Figuras 5.16 e 5.17), observa-se que durante o *fine-tuning* houve uma rápida convergência das perdas de treinamento e validação já nas primeiras épocas, sem indicativos claros de *over-fitting*. A maior melhoria das métricas ocorre nas duas primeiras épocas, estabilizando-se posteriormente. Com isso, as cinco épocas estabelecidas se mostraram ideais para maximizar o desempenho sem custos computacionais excessivos.

Figura 5.16 Convergência da Perda de Treinamento e Validação durante o *Fine-Tuning* do "DeepSeek-R1-Distill-Llama-8B"

Fonte: Elaboração própria.

Figura 5.17 Evolução das Métricas de Validação (Acurácia, AUC, F1, Precisão e *Recall*) ao Longo das Épocas para o "DeepSeek-R1-Distill-Llama-8B"



Fonte: Elaboração própria.

Após o treinamento, o modelo "DeepSeek-R1-Distill-Llama-8B" passou por uma validação em um cenário desbalanceado (98% dos dados na classe 0 e 2% na classe 1), isso é evidenciado na Tabela 5.6. A tabela mostra o desempenho que o modelo identifica quase

todos os positivos, 749 acertos e 51 falsos-negativos, ao preço de muitos falsos-positivos (FP = 2.364), o Recall fica em 0,9363, porém a Precisão vai para 0,2406 e, consequentemente, o F1 = 0,3828. A Acurácia é de 0,9396, todavia, a acurácia não é o melhor critério em virtude do forte desbalanceamento. Em compensação, o AUC de 0,9822 indica uma excelente capacidade discriminativa do modelo.

A análise da matriz de custo revelou que o modelo mantém um custo aceitável quando falsos negativos são fortemente penalizados (0,06548 para FP=1/FN=5), embora esse custo aumente significativamente quando penalizam-se falsos positivos (0,29678 para FP=5/FN=1), refletindo a dificuldade natural imposta pelo desequilíbrio dos dados.

Tabela 5.6 Desempenho de Validação do Modelo em Amostra Desbalanceada (98% Classe 0, 2% Classe 1)

Modelo	Amostra			Métricas de Desempenho					Matriz de Confusão				Matriz de Custo	
	T. Total	Label 1	Label 0	Acurácia	Precisão	*Recall*	F1	AUC	TN	FP	FN	TP	FP=1 / FN=5	FP=5 / FN=1
deepseek-ai/DeepSeek-R1-Distill-Llama-8B	40.000	800	39.200	0,9396	0,2406	0,9363	0,3828	0,9822	36.836	2.364	51	749	0,065475	0,296775

Fonte: Elaboração própria.

De modo geral, a aplicação do modelo na validação com os dados desbalanceados demonstrou uma ligeira queda de performance, todavia, essa queda na validação não indica sobreajuste, mas sim o efeito estatístico da mudança de prevalência: com só 2% de positivos, qualquer erro nas predições positivas derruba rapidamente as métricas.

Capítulo 6

CONSIDERAÇÕES FINAIS

O substancial volume de compras públicas no Brasil, superior a R$76 bilhões movimentados em 2023, e a frequência de práticas ilícitas, como superfaturamento e conluios entre fornecedores, tornam imperativa a adoção de instrumentos tecnológicos que proporcionem suporte aos órgãos de controle na triagem de transações suspeitas. A complexidade técnico-jurídica dos extratos de aquisições dificulta a auditoria manual e, paralelamente, a participação de empresas já sancionadas pela Controladoria-Geral da União aumenta o risco de repetição de fraudes. Nesse cenário, modelos de Processamento de Linguagem Natural (PLN) emergem como uma via promissora, ao integrar a capacidade de análise de grandes bases textuais com a agilidade demandada pelos processos de fiscalização.

Partindo desse desafio, este trabalho comparou abordagens *zero-shot*, *few-shot* e *fine-tuning* na detecção automática de fornecedores previamente punidos, tomando como base os extratos de compras disponibilizados pelo portal de Dados Abertos. Os resultados demonstram que o ajuste supervisionado é decisivo: o "DeepSeek-R1-Distill-Llama-8B", refinado em cinco épocas, atingiu 95% de acurácia e F1-Score, com AUC de 98%, superando os demais modelos internos (Tabela 5.5). Além disso, a matriz de custo normalizada indica o menor impacto dos erros, ainda que os falsos positivos tenham recebido penalização cinco vezes superior à dos falsos negativos, evidenciando um equilíbrio robusto entre sensibilidade e especificidade. Esse modelo treinado encontra-se disponível em: https://huggingface.co/CleitonOERocha/deepseek-r1-distill-llama-8B-finetuned-nfe-detection

Esses resultados confirmam a viabilidade do emprego de modelos de PLN como pré-etapa de investigação automática: o sistema reduz o universo de compras a ser revisado por especialistas, mitigando a carga de trabalho humana e acelerando respostas institucionais. A robustez observada se apoia na convergência estável das curvas de *loss* ao longo do treinamento e na evolução consistente das métricas de validação, que já superam a marca de 95% a partir da terceira época e se mantêm praticamente inalteradas até a quinta iteração, um indicativo de que o regime de treinamento adotado atinge um ponto de saturação eficiente.

Apesar dos progressos alcançados, persistem certas limitações. A ausência de padronização adequada nos campos textuais dos portais oficiais requer etapas complexas de pré-processamento, as quais nem sempre são capazes de corrigir todas as inconsistências. Ademais, o enfoque restringiu-se aos dados da esfera federal, deixando exposta a necessidade de exploração das compras estaduais e municipais. Finalmente, a manutenção de modelos de larga escala demanda infraestrutura computacional robusta e atualização contínua das bases de conhecimento, aspectos que podem ser onerosos para instituições com recursos limitados.

Para pesquisas futuras, recomenda-se expandir a coleta a outras esferas governamentais e a fontes externas, por exemplo, bases da Receita Federal ou do Tribunal Superior Eleitoral, enriquecendo o conjunto de variáveis analisadas. Sugere-se também integrar técnicas de detecção de anomalias em larga escala, combinando modelos de linguagem com grafos de conhecimento e métodos de análise de redes, a fim de mapear relações entre empresas, sócios e agentes públicos. Finalmente, uma atualização quase em tempo real, alimentada diariamente, tende a aumentar a eficácia do sistema e permitir um monitoramento preventivo mais ágil contra fraudes em aquisições governamentais.

REFERÊNCIAS BIBLIOGRÁFICAS

ALAM, S.; CARTLEDGE, C. L.; NELSON, M. L. *Support for Various HTTP Methods on the Web*. arXiv, 2014. ArXiv:1405.2330 [cs]. Disponível em: <http://arxiv.org/abs/1405.2330>.

ALDANA, A.; FALCóN-CORTéS, A.; LARRALDE, H. *A machine learning model to identify corruption in M\'exico's public procurement contracts*. arXiv, 2022. ArXiv:2211.01478 [cs]. Disponível em: <http://arxiv.org/abs/2211.01478>.

BASDEVANT, O. et al. Assessing Vulnerabilities to Corruption in Public Procurement and Their Price Impact. *IMF Working Papers*, v. 2022, n. 094, p. 1, maio 2022. ISSN 1018-5941. Disponível em: <https://elibrary.imf.org/openurl?genre=journal&issn=1018-5941&volume=2022&issue=094>.

BRAZ, C. S. et al. Exploring Irregularities in Brazilian Public Bids: An In-depth Analysis on Small Companies. *Journal on Interactive Systems*, v. 15, n. 1, p. 349–361, abr. 2024. ISSN 2763-7719. Number: 1. Disponível em: <https://journals-sol.sbc.org.br/index.php/jis/article/view/3836>.

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

CHEN, H.-J.; HUANG, S.-Y.; KUO, C.-L. Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets. *Expert Systems with Applications*, v. 36, n. 2, Part 1, p. 1478–1484, mar. 2009. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417407006045>.

COSTA, L. L. et al. Alertas de fraude em licitações: Uma abordagem baseada em redes sociais. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. SBC, 2022. p. 37–48. ISSN: 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/20515>.

COSTA, R. E.; HOLLNAGEL, H. C.; BUENO, R. L. P. COMPRAS GOVERNAMENTAIS: PANORAMA ATUAL E DESAFIOS GOVERNMENT PURCHASES: CURRENT OVERVIEW AND CHALLENGES. n. 23, 2019.

DEEPSEEK-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv, 2025. ArXiv:2501.12948 [cs]. Disponível em: <http://arxiv.org/abs/2501.12948>.

DEVLIN, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv, 2018. ArXiv:1810.04805 [cs]. Disponível em: <http://arxiv.org/abs/1810.04805>.

DOMINGOS, S. L. et al. Identifying IT Purchases Anomalies in the Brazilian Government Procurement System Using Deep Learning. In: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Anaheim, CA: IEEE, 2016. p. 722–727. ISBN 978-1-5090-6167-9. Disponível em: <https://ieeexplore.ieee.org/document/7838233/>.

FREITAS, M. d.; MALDONADO, J. M. S. d. V. O pregão eletrônico e as contratações de serviços contínuos. *Revista de Administração Pública*, v. 47, p. 1265–1281, out. 2013. ISSN 0034-7612, 1982-3134. Publisher: Fundação Getulio Vargas. Disponível em: <https://www.scielo.br/j/rap/a/jN7nbYZsHmtf8NvYrrd34jh/?format=html&lang=pt>.

HE, P. et al. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. arXiv, 2021. ArXiv:2006.03654 [cs]. Disponível em: <http://arxiv.org/abs/2006.03654>.

HOFFMANN, J. et al. *Training Compute-Optimal Large Language Models*. arXiv, 2022. ArXiv:2203.15556 [cs]. Disponível em: <http://arxiv.org/abs/2203.15556>.

HOTT, H. R. et al. Evaluating Contextualized Embeddings for Topic Modeling in Public Bidding Domain. In: NALDI, M. C.; BIANCHI, R. A. C. (Ed.). *Intelligent Systems*. Cham: Springer Nature Switzerland, 2023. p. 410–426. ISBN 978-3-031-45392-2.

HOULSBY, N. et al. *Parameter-Efficient Transfer Learning for NLP*. arXiv, 2019. ArXiv:1902.00751 [cs, stat]. Disponível em: <http://arxiv.org/abs/1902.00751>.

HOWARD, J.; RUDER, S. *Universal Language Model Fine-tuning for Text Classification*. arXiv, 2018. ArXiv:1801.06146 [cs, stat]. Disponível em: <http://arxiv.org/abs/1801.06146>.

ISHIKAWA, L.; ALENCAR, A. C. d. Compliance inteligente: o uso da inteligência artificial na integridade das contratações públicas. *Revista de Informação Legislativa*, v. 57, n. 225, p. 83–98, 2020. ISSN 0034-835x. Publisher: Senado Federal. Disponível em: <https://www12.senado.leg.br/ril/edicoes/57/225/ril_v57_n225_p83>.

JIANG, A. Q. et al. *Mistral 7B*. arXiv, 2023. ArXiv:2310.06825 [cs]. Disponível em: <http://arxiv.org/abs/2310.06825>.

KASHAP, S. PUBLIC PROCUREMENT AS A SOCIAL, ECONOMIC AND POLITICAL POLICY. In: . [s.n.], 2004. Disponível em: <https://www.semanticscholar.org/paper/PUBLIC-PROCUREMENT-AS-A-SOCIAL\%2C-ECONOMIC-AND-POLICY-Kashap/ff592e3f60f9f6c612155bafefae828c3b6dc1c2>.

LIMA, M. et al. Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In: COHN, T.; HE, Y.; LIU, Y.

(Ed.). *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 1580–1588. Disponível em: <https://aclanthology.org/2020.findings-emnlp.143>.

LUCCIONI, A. S.; VIGUIER, S.; LIGOZAT, A.-L. *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model*. arXiv, 2022. ArXiv:2211.02001 [cs]. Disponível em: <http://arxiv.org/abs/2211.02001>.

MCCANN, B. et al. *Learned in Translation: Contextualized Word Vectors*. arXiv, 2018. ArXiv:1708.00107 [cs]. Disponível em: <http://arxiv.org/abs/1708.00107>.

MEIRELLES, H. L. Licitação e contrato administrativo: de acordo com a Lei 8.666, de 21.6.1993, com todas as alterações posteriores. 2010. Accepted: 2010-08-04T17:33:49Z Publisher: Malheiros. Disponível em: <https://bdjur.stj.jus.br/jspui/handle/2011/32310>.

MOHALLEM, M. F.; RAGAZZO, C. E. J. Diagnóstico institucional: primeiros passos para um plano nacional anticorrupção. abr. 2017. Publisher: FGV Direito Rio. Disponível em: <https://hdl.handle.net/10438/18167>.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: *Sistemas Inteligentes Fundamentos e Aplicações*. 1. ed. Barueri-SP: Manole Ltda, 2003. p. 89–114. ISBN 85-204-168.

MOSBACH, M.; ANDRIUSHCHENKO, M.; KLAKOW, D. ON THE STABILITY OF FINE-TUNING BERT: MISCON- CEPTIONS, EXPLANATIONS, AND STRONG BASELINES. 2021.

NAI, R.; SULIS, E.; MEO, R. Public Procurement Fraud Detection and Artificial Intelligence Techniques: a Literature Review. 2022.

OECD. *Bribery in Public Procurement: Methods, Actors and Counter-Measures*. OECD, 2007. ISBN 978-92-64-01394-0 978-92-64-01396-4. Disponível em: <https://www.oecd-ilibrary.org/governance/bribery-in-public-procurement_9789264013964-en>.

ORMEROD, C. M.; PATEL, M.; WANG, H. *Using Language Models to Detect Alarming Student Responses*. arXiv, 2023. ArXiv:2305.07709 [cs]. Disponível em: <http://arxiv.org/abs/2305.07709>.

PARK, S. H.; GOO, J. M.; JO, C.-H. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean Journal of Radiology*, v. 5, n. 1, p. 11–18, 2004. ISSN 1229-6929. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698108/>.

PARNAMI, A.; LEE, M. *Learning from Few Examples: A Summary of Approaches to Few-Shot Learning*. arXiv, 2022. ArXiv:2203.04291 [cs]. Disponível em: <http://arxiv.org/abs/2203.04291>.

PETERS, M. E. et al. *Deep contextualized word representations*. arXiv, 2018. ArXiv:1802.05365 [cs]. Disponível em: <http://arxiv.org/abs/1802.05365>.

PLAčEK, M. et al. Analysis of Factors of Overpricing in Public Procurement: A Study for Low-performing EU Countries. *International Journal of Public Administration*, v. 43, n. 4, p. 350–360, mar. 2020. ISSN 0190-0692. Publisher: Routledge _eprint: https://doi.org/10.1080/01900692.2019.1636393. Disponível em: <https://doi.org/10.1080/01900692.2019.1636393>.

PRATI, R. C.; BATISTA, G. E. d. A. P. A.; MONARD, M. C. Uma experiência no balanceamento artificial de conjuntos de dados para aprendizado com classes desbalanceadas utilizando análise roc. (cdrom). In: *Jornadas Chilenas de Computacion*. [S.l.]: Universidade del BIO-BIO/Sociedade Chilena de Ciencia de la Computacion, 2003.

RADFORD, A. et al. Improving Language Understanding by Generative Pre-Training. 2018.

RIBEIRO, C. G. et al. Unveiling the public procurement market in Brazil: A methodological tool to measure its size and potential. *Development Policy Review*, v. 36, n. S1, p. O360–O377, 2018. ISSN 1467-7679. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/dpr.12301. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/dpr.12301>.

RIBEIRO, C. G.; JúNIOR, E. I. *O mercado de compras governamentais brasileiro (2006-2017): Mensuração e análise*. [S.l.], 2019. Disponível em: <https://www.econstor.eu/bitstream/10419/211431/1/1669548422.pdf>.

SILVA, M. O. et al. Overpricing Analysis in Brazilian Public Bidding Items. *Journal on Interactive Systems*, v. 15, n. 1, p. 130–142, jan. 2024. ISSN 2763-7719. Number: 1. Disponível em: <https://journals-sol.sbc.org.br/index.php/jis/article/view/3831>.

SIVARAJKUMAR, S.; WANG, Y. HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing. *AMIA Annual Symposium Proceedings*, v. 2022, p. 972–981, abr. 2023. ISSN 1942-597X. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10148337/>.

SKIENA, S. S. *The Data Science Design Manual*. Cham: Springer International Publishing, 2017. (Texts in Computer Science). ISBN 978-3-319-55443-3 978-3-319-55444-0. Disponível em: <http://link.springer.com/10.1007/978-3-319-55444-0>.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8.

SUN, C. et al. *How to Fine-Tune BERT for Text Classification?* arXiv, 2020. ArXiv:1905.05583 [cs]. Disponível em: <http://arxiv.org/abs/1905.05583>.

SUN, T.; SALES, L. J. Predicting Public Procurement Irregularity: An Application of Neural Networks. *Journal of Emerging Technologies in Accounting*, v. 15, n. 1, p. 141–154, jul. 2018. ISSN 1554-1908. Disponível em: <https://doi.org/10.2308/jeta-52086>.

TEAM, G. et al. *Gemma: Open Models Based on Gemini Research and Technology*. arXiv, 2024. ArXiv:2403.08295 [cs]. Disponível em: <http://arxiv.org/abs/2403.08295>.

TORRES-BERRU, Y.; BATISTA, V. F. L. Data Mining to Identify Anomalies in Public Procurement Rating Parameters. *Electronics*, v. 10, n. 22, p. 2873, jan. 2021. ISSN 2079-9292. Number: 22 Publisher: Multidisciplinary Digital Publishing Institute. Disponível em: <https://www.mdpi.com/2079-9292/10/22/2873>.

TORRES-BERRU, Y.; LOPEZ-BATISTA, V.; ZHINGRE, L. A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement. *Intelligent Automation & Soft Computing*, v. 36, n. 3, p. 3501–3516, 2023. ISSN 1079-8587, 2326-005X. Publisher: Tech Science Press. Disponível em: <https://www.techscience.com/iasc/v36n3/51926>.

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

TRANSPARÊNCIA, P. da. *Notas Fiscais*. 2024. <https://portaldatransparencia.gov.br/notas-fiscais>. Acessado: 23 de julho de 2024.

VASWANI, A. et al. *Attention Is All You Need*. arXiv, 2017. ArXiv:1706.03762 [cs]. Disponível em: <http://arxiv.org/abs/1706.03762>.

WANG, A. et al. *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. arXiv, 2019. ArXiv:1804.07461 [cs]. Disponível em: <http://arxiv.org/abs/1804.07461>.

WANG, H. Utilizing Imbalanced Data and Classification Cost Matrix to Predict Movie Preferences. *International Journal of Artificial Intelligence & Applications*, v. 9, n. 6, p. 01–12, nov. 2018. ISSN 09762191, 0975900X. ArXiv:1812.02529 [cs]. Disponível em: <http://arxiv.org/abs/1812.02529>.

WANG, Z.; PANG, Y.; LIN, Y. *Large Language Models Are Zero-Shot Text Classifiers*. arXiv, 2023. ArXiv:2312.01044 [cs]. Disponível em: <http://arxiv.org/abs/2312.01044>.

WOLF, T. et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv, 2020. ArXiv:1910.03771 [cs]. Disponível em: <http://arxiv.org/abs/1910.03771>.