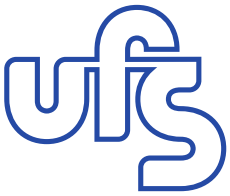


UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Análise Exploratória e Experimental de Aplicações de Inteligência Artificial para Classificação de Descrições Incongruentes em Compras na Área de Saúde Pública

Dissertação de Mestrado

Wesckley Faria Gomes

São Cristóvão – Sergipe

2024

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Wesckley Faria Gomes

Análise Exploratória e Experimental de Aplicações de Inteligência Artificial para Classificação de Descrições Incongruentes em Compras na Área de Saúde Pública

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Prof. Dr. Methanias Colaço Rodrigues Júnior

São Cristóvão – Sergipe

2024

À minha família, pelo amor incondicional, apoio constante e incentivo ao longo dos anos. Sem vocês, essa conquista não seria possível.

Agradecimentos

Primeiramente à Deus, dando-me sabedoria, força e serenidade para enfrentar e vencer os obstáculos que cruzam minha caminhada.

Ao meu pai Léo e meu irmão Wesley, por todo o apoio e por estarem sempre ao meu lado.

A minha amada esposa Aline pela compreensão, apoio e paciência.

Ao meu orientador, Prof. Dr. Methanias Colaço Rodrigues Júnior, pela paciência, pelo apoio, pelo incentivo, pela sua disponibilidade total e, principalmente, pelos valiosos conselhos e orientações.

Aos amigos e companheiros de mestrado que contribuíram diretamente nesta minha formação.

Obrigado a todos que, de alguma maneira, contribuíram para o meu crescimento pessoal e para a realização deste trabalho.

E, acima de tudo, dedico este trabalho à memória da minha mãe Elza, cujo amor e inspiração continuam a me guiar, mesmo ausentes fisicamente.

"A imaginação é mais importante que o conhecimento."
(Albert Einstein)

Resumo

Contexto: O setor de Órteses, Próteses e Materiais Especiais (OPME) na área da saúde apresenta uma ampla variedade de produtos e tecnologias, envolvendo tanto empresas multinacionais quanto locais. Apesar dos avanços tecnológicos, muitos serviços e sistemas de informação, especialmente no setor público, ainda utilizam descrições não estruturadas em linguagem natural de produtos, serviços ou eventos, dificultando suas classificações e análises. Todavia, para auditorias eficientes, é necessário classificar e totalizar automaticamente faturas emitidas para compra de produtos. Desta forma, a falta de padronização na nomenclatura na comercialização de OPMEs, não apenas dificulta a comparação dos produtos, seja para uniformização de preços ou padronização de uso, mas também abre espaço para possíveis atos de corrupção. **Objetivo:** Para mitigar o problema de padronização e codificação ineficazes, desenvolver e avaliar a eficácia e eficiência de um classificador de OPMEs, no contexto de descrições de notas fiscais eletrônicas, do ponto de vista de auditores, profissionais de saúde e cientistas de dados. **Método:** Inicialmente, foi realizado um Mapeamento Sistemático (MS), como forma de identificar e caracterizar as abordagens e técnicas de inteligência artificial para a classificação automática de descrições textuais incongruentes em notas fiscais. Em seguida, foi implementada uma ferramenta baseada em inteligência artificial, o **OPM**in**E**r, para classificar notas fiscais de OPMEs. Ato contínuo, foi realizado um experimento controlado para avaliar os algoritmos de Inteligência Artificial (IA) mapeados. **Resultados:** A estratégia de busca utilizada no mapeamento sistemático selecionou 225 artigos, os quais passaram pelos critérios de inclusão e exclusão. Dentre as abordagens encontradas para resolução do problema de descrições textuais incongruentes, o destaque foi para o aprendizado de máquina supervisionado, presente em 60% dos trabalhos. Já no experimento controlado, considerando a significância estatística, o algoritmo *Linear Support Vector* alcançou uma acurácia de 99% e se destacou dentre os demais. Em termos de eficiência, o algoritmo *Naïve Bayes Multinomial* se destacou, tendo o tempo médio de treinamento mais rápido, com 4,375 segundos. **Conclusão:** Os resultados mostraram que é possível identificar e classificar OPMEs em notas fiscais de forma automática. Isso permite uma análise mais precisa e eficaz de indícios tais como, por exemplo, preços anomalamente altos e quantidades de OPMEs compradas por habitante, os quais são analisados pela Auditoria do Sistema Único de Saúde (AudSUS), Ministério da Saúde - Brasil, para identificação de potenciais irregularidades e contribuição para a transparência e eficiência na gestão de recursos da área da saúde.

Palavras-chave: Órteses, Próteses e Materiais Especiais (OPME), saúde, auditoria, classificação, inteligência artificial (IA), notas fiscais eletrônicas, corrupção.

Abstract

Context: The Orthoses, Prostheses and Special Materials (OPME) sector in the health area presents a wide variety of products and technologies, involving both multinational companies as local. Despite technological advances, many services and information systems, especially in the public sector, still use unstructured descriptions in natural language of products, services or events, making classification and analysis difficult. However, for efficient audits, it is necessary to automatically classify and total invoices issued for purchasing products. In this way, the lack of standardization in the nomenclature in commercialization of OPMEs, not only makes it difficult to compare products, either for standardizing prices or standardization of use, but it also opens up space for possible acts of corruption. **Objective:** To mitigate the problem of ineffective standardization and coding, develop and evaluate the effectiveness and efficiency of an OPME classifier, in the context of electronic invoice descriptions, from the perspective of auditors, health professionals and data scientists. **Method:** Initially, Systematic Mapping (SM) was carried out as a way to identify and characterize the Artificial intelligence approaches and techniques for automatically classifying incongruent textual descriptions on invoices. Subsequently, an artificial intelligence-based tool, OPMinEr, was implemented to classify OPME invoices. A controlled experiment was then carried out to evaluate the mapped Artificial Intelligence (AI) algorithms. **Results:** The search strategy used in the systematic mapping selected 225 articles, which passed the inclusion and exclusion criteria. Among the approaches found to solve the problem of incongruent textual descriptions, supervised machine learning was prominent in 60% of the works. In the controlled experiment, considering statistical significance, the Linear Support Vector algorithm achieved an accuracy of 99% and stood out among the others. In terms of efficiency, the algorithm Naïve Bayes Multinomial stood out, having the fastest average training time, with 4.375 seconds. **Conclusion:** The results demonstrate that it is possible to automatically identify and classify OPMEs in invoices, allowing for a more precise analysis of indicators such as anomalously high prices and quantities of OPMEs purchased per inhabitant. These analyses are conducted by the Audit of the Unified Health System (AudSUS), Ministry of Health - Brazil, to identify potential irregularities and contribute to transparency and efficiency in healthcare resource management.

Keywords: Orthoses, Prostheses and Special Materials (OPME), health, audit, classification, artificial intelligence (AI), electronic invoices, corruption.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ABIIS	Aliança Brasileira da Indústria Inovadora em Saúde
ANS	Agência Nacional de Saúde
ANVISA	Agência Nacional de Vigilância Sanitária
AudSUS	Auditoria do Sistema Único de Saúde
BERT	Bidirectional Encoder Representations from Transformers
CATMAT	Catálogo de Materiais
CGU	Controladoria Geral da União
CNN	Redes Neurais Convulacionais
DANFE	Documento Auxiliar da Nota Fiscal Eletrônica
FIESP	Federação das Indústrias do Estado de São Paulo
FN	False Negative
FP	False Positive
GQM	Goal Question Metric
HS	Harmonized System
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IBTP	Instituto Brasileiro de Planejamento e Tributação
ICMS	Imposto sobre Circulação de Mercadorias e Prestação de Serviços
IDF	Inverse Document-Frequency
MS	Mapeamento Sistemático
NCM	Nomenclatura Comum do Mercosul
ONG	Organização Não Governamental
OPME	Órteses, Próteses e Materiais Especiais
RNN	Redes Neurais Recorrentes

SH	Sistema Harmonizado
TF	Term-Frequency
TN	True Negative
TP	True Positive

Sumário

1 Introdução

Este capítulo tem o propósito de estabelecer um contexto conciso relacionado ao tema da pesquisa, explicar a motivação por trás do estudo, destacar a problemática que será abordada, apresentar as questões de pesquisa que direcionarão o estudo, esclarecer os objetivos a serem alcançados e propor hipóteses a serem testadas. Além disso, serão delineadas as contribuições esperadas ao final do trabalho, juntamente com uma visão geral da metodologia de pesquisa a ser adotada.

1.1 Contextualização

A corrupção é caracterizada como a utilização indevida do poder com o intuito de obter vantagens ilegais, e pode ser identificada em praticamente todas as formas de organizações ou grupos, abrangendo desde instituições governamentais até empresas privadas e organizações sem fins lucrativos (KRATCOSKI, 2018). A Convenção contra a Corrupção da Organização das Nações Unidas (ONU) descreve a corrupção como uma praga insidiosa, que tem uma ampla gama de efeitos corrosivos nas sociedades: mina a democracia e o estado de direito, leva a violações dos direitos humanos, distorce os mercados, prejudica a qualidade de vida e permite que a organização do crime, o terrorismo e outras ameaças à segurança humana floresçam (NATIONS, 2004).

A quantificação da corrupção é uma tarefa desafiadora, uma vez que suas dimensões, componentes e envolvidos, muitas vezes operam de forma oculta. Devido a esses desafios, os pesquisadores e organizações usam uma variedade de abordagens para medir a corrupção, incluindo índices de percepção, dados financeiros, pesquisas de campo e análises qualitativas. Em resposta a essa complexidade, a avaliação da corrupção é frequentemente realizada de maneira indireta, centrando-se na medição da percepção desse fenômeno pela população. Nesse contexto, destaca-se o Índice de Percepção da Corrupção, elaborado pela Organização Não Governamental

(ONG) Transparência Internacional, que é amplamente reconhecida. Esse índice utiliza uma abordagem abrangente que se baseia em treze fontes distintas de dados. Isso inclui pesquisas realizadas com especialistas no campo da corrupção e executivos de diversos setores da economia, juntamente com indicadores socioeconômicos quantitativos e qualitativos. A integração dessas diversas fontes de dados possibilita a comparação dos níveis de percepção da corrupção entre diferentes países (INTERNATIONAL, 2022).

A medição da corrupção e seus impactos na economia também são calculados usando modelos teóricos. No Brasil, a Federação das Indústrias do Estado de São Paulo (FIESP) realizou um estudo a esse respeito em 2010 (FIESP, 2010). Utilizando um modelo de crescimento econômico neoclássico baseado na pesquisa de Mankiw et al. (1992), estimou-se que as perdas do PIB devidas à corrupção poderiam variar entre 1,38% e 2,30%. Em termos do PIB brasileiro de 2019, isso equivale a um montante entre R$ 100 bilhões e R$ 160 bilhões, de acordo com os cálculos do Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2020).

No cenário das possibilidades de corrupção, os gastos governamentais desempenham um papel significativo na economia do Brasil. No exercício de 2020, por exemplo, o orçamento federal previa a alocação de quatro trilhões de reais, conforme informações da Controladoria Geral da União (CGU) (CGU, 2020). Tais recursos são, de forma recorrente, alvo de ações corruptas no Brasil, notadamente aqueles ligados às áreas de saúde e educação públicas. A Constituição Federal estabelece que os estados devem destinar, no mínimo, 12% de sua receita para a área da saúde; que os municípios devem aplicar pelo menos 15% de seus recursos nessa seara, e a União deve destinar 15% de sua Receita Corrente Líquida para financiar o setor (BRASIL, 1988) (VIEIRA F. S.; PIOLA, 2019). Essas disposições legais, embora destinadas a garantir o financiamento adequado de serviços essenciais, também podem criar oportunidades para a corrupção, uma vez que os recursos consideráveis alocados para essas áreas podem ser desviados ou mal utilizados. A gestão transparente e eficaz dos gastos governamentais é crucial para garantir que esses recursos sejam direcionados para os fins pretendidos, beneficiando a população e promovendo o desenvolvimento socioeconômico.

Os recursos financeiros do governo federal destinados a despesas relacionadas à saúde são transferidos para os Estados, o Distrito Federal e os Municípios por meio de dois tipos de financiamento: o Bloco de Custeio das Ações e Serviços Públicos de Saúde e o Bloco de Investimento na Rede de Serviços Públicos de Saúde. Cada um desses blocos é subdividido em categorias específicas, e todo o sistema contábil abrange uma ampla variedade de programas e ações orçamentárias relacionadas à área da saúde. Isso inclui diversas iniciativas que envolvem a compra de Órteses, Próteses e Materiais Especiais (OPME) pelo Ministério da Saúde, conforme estabelecido pela Portaria n. 3.992 (MS, 2017). Segundo a Aliança Brasileira da Indústria Inovadora em Saúde (ABIIS) (ABIIS, 2022), as importações com Dispositivos Médicos, incluindo OPMEs, foi de U$ 6,3 bilhões no acumulado de janeiro a dezembro de 2022, tendo os Estados Unidos o maior fornecedor. Hodiernamente, os entes governamentais devem manter portais da

transparência que evidenciam a arrecadação de receitas e as despesas executadas diariamente, porém, de acordo com Paiva & Revoredo (2016), a mera disponibilização dessas informações em portais governamentais não assegura um aumento efetivo do grau de transparência desses entes, pois o grande volume de dados aliado à falta de padrões torna inviável qualquer tipo de acompanhamento sistemático desses dados. Portanto, faz-se necessário o desenvolvimento de ferramentas mais amigáveis para auxiliar o cidadão na análise de dados e auditoria dos gastos públicos (CORREA; LEAL, 2018).

Em se tratando da área de Saúde, o Mercado de Órteses, Próteses e Materiais Especiais (OPME) é marcado pela quantidade de produtos existentes e pela diversidade das tecnologias utilizadas. É um setor com o domínio de empresas multinacionais, porém também conta com pequenas e médias empresas locais, o que causa uma heterogeneidade desses dispositivos, concentrando o conhecimento em especialistas e produzindo assimetria de informação (CRUZ et al., 2022). Isto, alinhado com a dificuldade de padronização da nomenclatura na comercialização de OPMEs, dificulta a comparação dos produtos, seja para uniformização de preços ou padronização de uso, o que dá margem para atos de corrupção.

Em janeiro de 2015, a imprensa nacional noticiou indícios da ocorrência de esquema fraudulento envolvendo a compra e utilização de órteses, próteses e materiais especiais (OPME), o que ficou conhecido como “máfia das próteses”. O suposto esquema envolveria uma série de atores – fabricantes, distribuidores, hospitais, médicos e advogados – e diversos tipos de irregularidades – venda de dispositivos com sobrepreço, recebimento de comissões irregulares, fraudes, desvios, entre outros (TCU, 2016).

1.2 Análise do Problema

Segundo Ribeiro et al. (2018), os emitentes de notas fiscais informam em linguagem natural a descrição dos seus produtos e utilizam uma linguagem comercial e comunicativa, sem a preocupação com a exatidão dos termos. Além disso, os itens de produtos descritos possuem códigos associados que correspondem ao imposto que incidirá sobre os produtos, porém, alguns contribuintes, agindo por falta de conhecimento ou intencionalmente, associam códigos que não correspondem de fato aos itens de produtos descritos, de modo que alíquotas distintas incidem sobre os mesmos, fazendo com que o Estado deixe de arrecadar o referido imposto (BATISTA; BAGATINI; FROZZA, 2018). No que se refere ao comércio internacional, na importação e exportação de mercadorias, o Sistema Harmonizado de Descrição e Codificação de Mercadorias, ou simplesmente Sistema Harmonizado (HS em inglês: Harmonized System) é utilizado para padronizar os produtos comercializados. De acordo com Ding et al. (2015), estudos comerciais mostram que 30% dos envios de declarações utilizam o código Harmonized System (HS) errado, ou seja, o código informado não corresponde à descrição informada.

A padronização com o código HS é feita a partir de um código, uma descrição e uma

unidade, que são utilizados como base para decidir a tarifa alfandegária. Na sua forma básica, o código é composto por 6 dígitos, subdividido hierarquicamente em grupos de 2 dígitos (capítulos, títulos e subtítulos). Segundo Chen et al. (2021), a tarefa de atribuir o código correto é feita por um especialista e é uma tarefa intensa e propícia a erros. Cada país pode utilizar seu próprio código. Por exemplo, os países que compõem o bloco econômico do Mercosul utilizam a Nomenclatura Comum do Mercosul (NCM), uma convenção de categorização de mercadorias adotada desde 1995, a qual é composta por oito dígitos, sendo os seis primeiros dígitos herdados do Sistema Harmonizado (SH), acrescidos de dois dígitos, que são específicos do âmbito do Mercosul (BATISTA; BAGATINI; FROZZA, 2018).

O grande volume de dados de notas fiscais emitidas e descrições de produtos e serviços em linguagem natural não estruturada dificulta o processo de auditoria e investigação de fraudes por parte dos órgãos competentes. Em 2017, o Tribunal de Contas Europeu comunicou que as formas de evasão de pagamento amplamente aplicadas são a subavaliação, classificação errada ao mudar para uma classificação de um produto com alíquota mais baixa, e a descrição errada das mercadorias importadas. No período de 2013 a 2016, cerca de 2 bilhões de euros foram perdidos em direitos aduaneiros devido à subvalorização das importações de têxteis e calçados da China para o Reino Unido (SPICHAKOVA; HAAV, 2020).

Diante desses problemas apresentados e suas limitações impostas, as seguintes questões de pesquisa foram elaboras.

Questão principal de pesquisa:

- **Q1:** Um classificador, baseado em inteligência artificial, pode automatizar a classificação de OPMEs em notas fiscais?

Questões secundárias de pesquisa:

- **Q2:** Entre os algoritmos selecionados, qual o melhor em termos de eficácia?
- **Q3:** Entre os algoritmos selecionados, qual o melhor em termos de eficiência?

Para responder a questão de pesquisa **Q1**, uma Prova de Conceito (PoC) foi elaborada (vide seção 3) para avaliar uma ferramenta na tarefa de identificação e classificação de notas fiscais de OPMEs. Para responder as questões secundárias de pesquisa **Q2** e **Q3**, especificamente, foi realizado um experimento controlado (vide seção 4), o qual elucidou, operacionalizou e testou as seguintes hipóteses teóricas nulas, seguida de suas hipóteses alternativas.

Formulação de Hipóteses:

Para responder à questão Q2, as seguintes hipóteses foram elaboradas:

- H_0**:** Os algoritmos (1, 2..n) têm a mesma eficácia.

- H_1**:** Os algoritmos (1, 2..n) têm eficácias diferentes.

Para responder à questão Q3, as seguintes hipóteses foram elaboradas:

- H_0**:** Os algoritmos (1, 2..n) têm a mesma eficiência.
- H_1**:** Os algoritmos (1, 2..n) têm eficiências diferentes.

1.3 Justificativa

A corrupção é uma ameaça significativa em todas as esferas da sociedade, e a área de saúde não está imune a práticas fraudulentas. A diversidade e a falta de padronização no mercado de OPMEs contribuem para assimetrias de informação, dificultando a comparação de produtos e abrindo espaço para atos de corrupção. A fragilidade na identificação de produtos nas notas fiscais, devido a erros na codificação e descrição, dificulta as atividades de fiscalização e auditoria.

O desenvolvimento de um classificador de textos, baseado em inteligência artificial, pode ser a solução para a correta correlação das descrições de OPMEs constantes nas NF-es às nomenclaturas NCM, servindo efetivamente aos trabalhos investigativos realizados pelos Órgãos de Controle. Tal utilidade é ilustrada pela identificação de fraudes no fornecimento destes produtos, como por exemplo, OPMEs vendidas com valores anomalamente altos, relação anômalas entre profissinoais solicitantes e autorizadores, produtos comprados próximo ou fora do vencimento.

1.4 Objetivos

O objetivo geral deste trabalho foi a proposta, implementação e avaliação de uma ferramenta, baseada em inteligência artificial, para identificar e classificar OPMEs em descrições de notas fiscais.

1.4.1 Objetivos Específicos

Os objetivos norteadores do presente trabalho são expostos a seguir:

1. Mapeamento sistemático da literatura, com a finalidade de analisar o estado da arte da inteligência artificial no ambiente de descrições textuais incongruentes em notas fiscais.
2. Classificador baseado em algoritmos utilizados na literatura.
3. Experimento controlado para avaliar a eficácia e eficiência do classificador, sob a perspectiva de diferentes medidas e diferentes algoritmos de classificação.

1.5 Metodologia

A metodologia adotada para o trabalho envolveu, inicialmente, um Mapeamento Sistemático (MS) da literatura, publicado em (GOMES; COLAÇO JÚNIOR, 2022), baseada no protocolo proposto por Kitchenham (2004), tendo por finalidade analisar o estado da arte da inteligência artificial no ambiente de descrições textuais incongruentes em notas fiscais.

Para o estudo, o primeiro passo foi a construção de um dicionário de dados, com descrições classificadas corretamente, que serviu como base de treinamento para os algoritmos. O modelo de conhecimento foi composto por registros oriundos da Agência Nacional de Vigilância Sanitária (Anvisa), cadastro oficial de materiais do governo federal (Catmat), descrições avulsas (inseridas manualmente) escritas por auditores e descrições das próprias notas fiscais.

Ato contínuo, foi desenvolvido o classificador **OPM**in**E**r, o qual foi avaliado preliminarmente mediante uma prova de conceito, cuja metodologia está detalhada no capítulo 3. Por fim, foi realizado um experimento controlado *"in vitro"*, envolvendo dados de notas fiscais de compras na área da saúde pública, detalhado no capítulo 4. De acordo com Wohlin et al. (2000), uma experimentação não é uma tarefa simples, pois envolve preparar, conduzir e analisar experimentos corretamente.

1.6 Organização da Dissertação

Este documento está organizado de acordo com a Instrução Normativa Nº 05/2019/PROCC, a qual permite que a Dissertação seja "uma compilação de artigos científicos submetidos ou publicados em veículos com Qualis". São 5 capítulos que fornecem uma base conceitual para o entendimento sistêmico. Os tópicos a seguir descrevem o conteúdo de cada um dos capítulos:

- O Capítulo 1 apresenta esta Introdução, explicando as justificativas juntamente com as hipóteses levantadas;
- O Capítulo 2 apresenta parte do artigo de Mapeamento Sistemático que foi publicado no Simpósio Brasileiro de Sistemas de Informação (SBSI 2022) (GOMES; COLAÇO JÚNIOR, 2022);
- O Capítulo 3 apresenta parte do artigo publicado no SBCAS 2023: XXIII Simpósio Brasileiro de Computação Aplicada à Saúde, que descreve uma prova de conceito do classificador (GOMES et al., 2023);
- O Capítulo 4 apresenta parte do artigo submetido à Revista Ciência & Saúde Coletiva, descrevendo um experimento controlado realizado com as descrições de notas fiscais;
- O Capítulo 5 apresenta a discussão do presente trabalho;

- Finalmente, no Capítulo 6, é apresentado um compilado de conclusões, contribuições e sugestões de trabalhos futuros.

2

Applications of Artificial Intelligence for Auditing and Classification of Incongruent Descriptions in Public Procurement

2.1 Planejamento do Mapeamento Sistemático

2.1.1 Objetivo

Identificar e caracterizar as abordagens e técnicas de inteligência artificial para a classificação automática de descrições textuais em notas fiscais.

2.1.2 Questões de Pesquisa

As questões de pesquisa foram desenvolvidas com o objetivo de apresentar uma visão geral da área, destacando aspectos fundamentais dos estudos primários (KITCHENHAM, 2004)(PETERSEN; VAKKALANKA; KUZNIARZ, 2015). Foram elaboradas tendo como base o modelo PICO (BERGIN; WRAIGHT, 2006)(SANTOS; PIMENTA; NOBRE, 2007), que tem como objetivo destacar os efeitos de uma intervenção em uma determinada população e estruturar a pesquisa em quatro elementos fundamentais: População, Intervenção, Controle e “Outcomes” (Resultados). Esses elementos, segundo Santos et al. (2007), podem ser utilizados para construir questões de pesquisa de naturezas diversas. A Tabela 1 ilustra o modelo PICO utilizado neste trabalho.

Assim, a partir da definição do modelo PICO, foram elaboradas questões de pesquisa, baseando-se nas diretrizes de protocolo de Mapeamento Sistemático da Literatura observados em (KITCHENHAM, 2004)(PETERSEN; VAKKALANKA; KUZNIARZ, 2015). São elas:

- Q1: Quais são as abordagens utilizadas para mitigar casos de descrições textuais incongruentes em notas fiscais?

Acrônimo	Categoria	Descrição
P	População	Publicações que abordem, de forma direta, a classificação de descrições textuais em notas fiscais.
I	Intervenção	Contexto de aplicações que utilizam abordagens, técnicas e algoritmos inteligentes para a classificação de descrições textuais em notas fiscais.
C	Controle	Aplicações que não utilizam algoritmos inteligentes para a classificação de descrições textuais em notas fiscais.
O	Resultados	Análise inteligente para a classificação de descrições textuais em notas fiscais.

Tabela 1 – Modelo PICO para conformidade das questões de pesquisa.

Fonte: Autores.

- Q2: Quais as principais técnicas utilizadas?
- Q3: Quais técnicas apresentam melhores eficácias?
- Q4: Quais os meios de publicações mais populares?
- Q5: Em que anos foram publicados mais artigos nesta área?
- Q6: Quais países têm mais publicações nesta área?

2.1.3 Estratégia de Busca e de Seleção

Para a execução da busca de artigos, foram selecionadas as bases de dados responsáveis pela publicação dos principais periódicos da área de Ciência da Computação. Foram elas: SCOPUS, ScienceDirect, IEEE, ACM e Spell. Para realização da busca, foram utilizadas as ferramentas de filtragem disponibilizadas em cada base de dados e foram considerados título, resumo e palavras-chave. O acesso às bases de dados foi realizado por meio do portal de periódicos da CAPES (CAPES, 2021), utilizando assinatura institucional, sem quaisquer restrições de artigos.

Para a realização da busca nas bases de dados digitais, foi definida uma *string* de busca composta por termos em inglês e sinônimos, associados à classificação de descrições textuais em notas fiscais. Os termos foram identificados a partir dos papéis do modelo PICO, definidos na Tabela 1, e posteriormente refinados e adaptados para melhor aproveitamento da *string*. A Tabela 2 apresenta os termos antes do refinamento.

Após o refinamento, os termos ajustados foram utilizados para construir a *string* de busca, os quais estão descritos na Tabela 3.

Categoria	**Descrição**
População	Invoice, Harmonized System, HS code, Government Bidding, Government Procurement, Goods Description, Government Purchase
Intervenção	Machine Learning, Data Mining, Text Classification, Text Mining, Artificial Intelligence, Deep Learning
Controle	-
Resultados	Classify, Fraud, Investigation, Identify, Misclassification, Analysis

Tabela 2 – Categorias do modelo PICO e termos identificados para pesquisa bibliográfica.

Fonte: Autores.

Termos da *string* de busca		
Invoice, Harmonized System, HS code, Government Bidding, Government Procurement, Goods Description, Government Purchase	Machine Learning, Data Mining, Text Classification, Text Mining, Artificial Intelligence, Deep Learning	Classify, Fraud, Investigation, Identify, Misclassification, Analysis

Tabela 3 – *Strings* eleitas após o refinamento.

Fonte: Autores.

A partir dos termos evidenciados acima, a seguinte *string* de busca foi elaborada:

- *TITLE-ABS-KEY (("Invoice"OR "Harmonized System"OR "HS Code"OR "Government Bidding"OR "Government Procurement"OR "Goods Description"OR "Government Purchas*") AND ("Machine Learning"OR "Data Mining"OR "Text Classification"OR "Text Mining"OR "Artificial Intelligence"OR "Deep Learning") AND ("Classif*"OR "Fraud"OR "Investigation"OR "Identif*"OR "Misclassification"OR "Analy*"))*

2.1.4 Critérios de Seleção de Fontes

Os critérios de inclusão e exclusão foram estabelecidos para filtrar os artigos relevantes para o mapeamento sistemático. Após a realização da busca, utilizando a *string* de busca apresentada na seção 2.1.3, foram contabilizados apenas os estudos selecionados para avaliação, os quais passaram pelos critérios de inclusão e exclusão. A seguir, são apresentados os critérios de inclusão e exclusão.

Critérios de Inclusão:

1. Artigos que estejam disponíveis de forma online e em bibliotecas digitais;
2. Artigos que contenham o texto da pesquisa no título, resumo ou palavras-chave;
3. Artigos devem explorar análise de algoritmo, técnicas, mecanismos ou abordagens de classificação de descrições textuais em notas fiscais;
4. Os artigos devem ter a data de publicação entre os anos de 2011 e 2021;
5. Artigos escritos em inglês.

Critérios de Exclusão:

1. Estudos duplicados;
2. *Surveys*;
3. Revisão Sistemática
4. Artigos com mais de dez anos de publicação;
5. Estudos preliminares;
6. Artigos Curtos (*Short Papers*).

2.1.5 Estratégia de Extração de Informações

Para avaliar a qualidade do trabalho e responder às questões de pesquisa expostas na seção 2.1.2, foi projetado um formulário a ser respondido para cada artigo lido na íntegra. Segundo Kitchenham (2004), os formulários de extração de dados devem ser projetados para coletar todas as informações necessárias para abordar as questões e os critérios de qualidade do estudo. A Tabela 4 apresenta o formulário de extração utilizado nesta pesquisa.

2.2 Condução do Mapeamento Sistemático

Responsável pela publicação dos principais periódicos na área de Ciência da Computação, a base de dados Scopus (SCOPUS, 2021) foi escolhida como base para definição e refinamento da *string* de busca. Esta base inclui também artigos de diferentes bases de dados científicas, tais como IEEE, ACM, Springer e Elsevier. Após definida, refinada e julgada adequada, a *string* foi traduzida para as outras máquinas de busca utilizadas neste trabalho, as quais foram: IEEE, ACM, Web of Science, ScienceDirect e Spell. No total, foram retornados 225 trabalhos, sendo 140 (62%) da Scopus, 36 (16%) da Web of Science, 20 (9%) do IEEE, 15 (7%) da ScienceDirect, 8 (3%) da ACM e 6 (3%) da Spell. Os dados estão representados na Figura 1.

1.	Qual a abordagem utilizada?	[Supervisionada, Não Supervisionada, Semissupervisionada, Sem Classificação, Apenas Pré-Processamento]
2.	Qual a técnica utilizada?	
3.	Qual o tipo de estudo utilizado?	[Aplicação Prática, Estudo de Caso, Prova de Conceito, Experimento Controlado]
4.	O estudo possui alguma avaliação experimental?	[Sim, Não]
5.	Foram declaradas as ameaças à validade?	[Sim, Não]

Tabela 4 – Formulário de Extração.

Fonte: Autores.

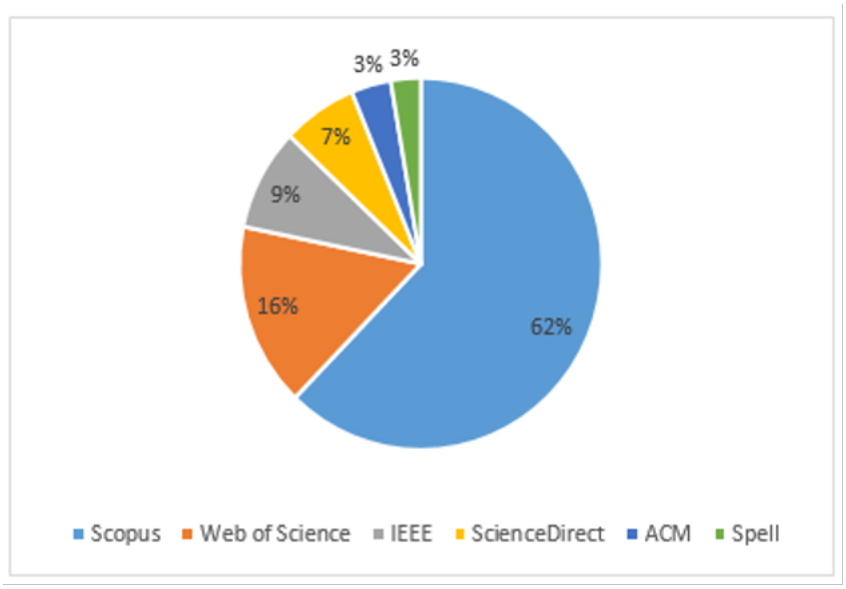


Figura 1 – Execução da *String* de Busca nas Bibliotecas Digitais.

Após a busca dos artigos nas bases de dados, iniciou-se o processo de filtragem, com base nos critérios de seleção definidos na seção 2.1.4. Cada artigo foi classificado como Aceito ou Rejeitado. Do total de 225 publicações analisadas, 76 (34% do total) não estavam de acordo com os critérios de exclusão ou estavam duplicadas. Após a remoção destes artigos, foi realizada a leitura superficial dos trabalhos restantes, analisando-se o título, resumo e palavras-chave. Ao final desta etapa, 115 (77% do total) artigos estavam fora do escopo deste mapeamento, sendo classificados como Rejeitados. Por fim, para análise detalhada, o restante dos artigos foi classificado como Aceito. A Figura 2 mostra o resumo desta etapa.

A fim de confirmar que os artigos restantes utilizavam abordagens de inteligência artificial

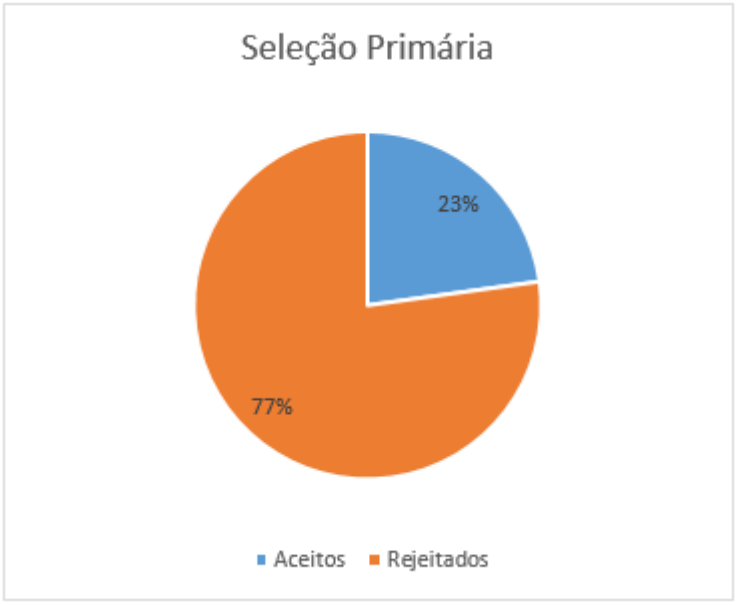


Figura 2 – Primeira Etapa de Seleção dos Trabalhos.

para classificação de descrições textuais em notas fiscais de compras, foi realizada a leitura completa dos textos. No entanto, não foi possível acessar todas as publicações, mesmo após enviar e-mail diretamente aos autores. Dos 34 artigos separados para avaliação, 32 (94%) foram recuperados e 2 (6%) não foi possível acessar o texto completo, os quais, consequentemente, foram classificados como rejeitados. Desta forma, 19 trabalhos estavam fora do escopo deste mapeamento e foram rejeitados. Como resultado final, 15 artigos foram aceitos. A Figura 3 mostra o resultado da seleção secundária.

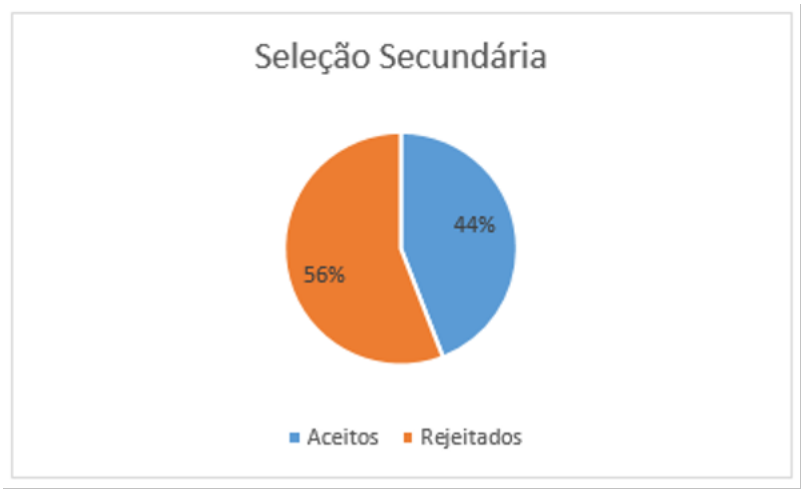


Figura 3 – Segunda Etapa de Seleção dos Trabalhos.

2.3 Síntese dos Dados e Apresentação dos Resultados

Nesta seção, serão apresentados os resultados do mapeamento sistemático. A Figura 4 apresenta um fluxo descrevendo o processo de extração dos artigos obtidos em cada etapa do processo e em seguida as questões de pesquisa são respondidas de acordo com os dados extraídos.

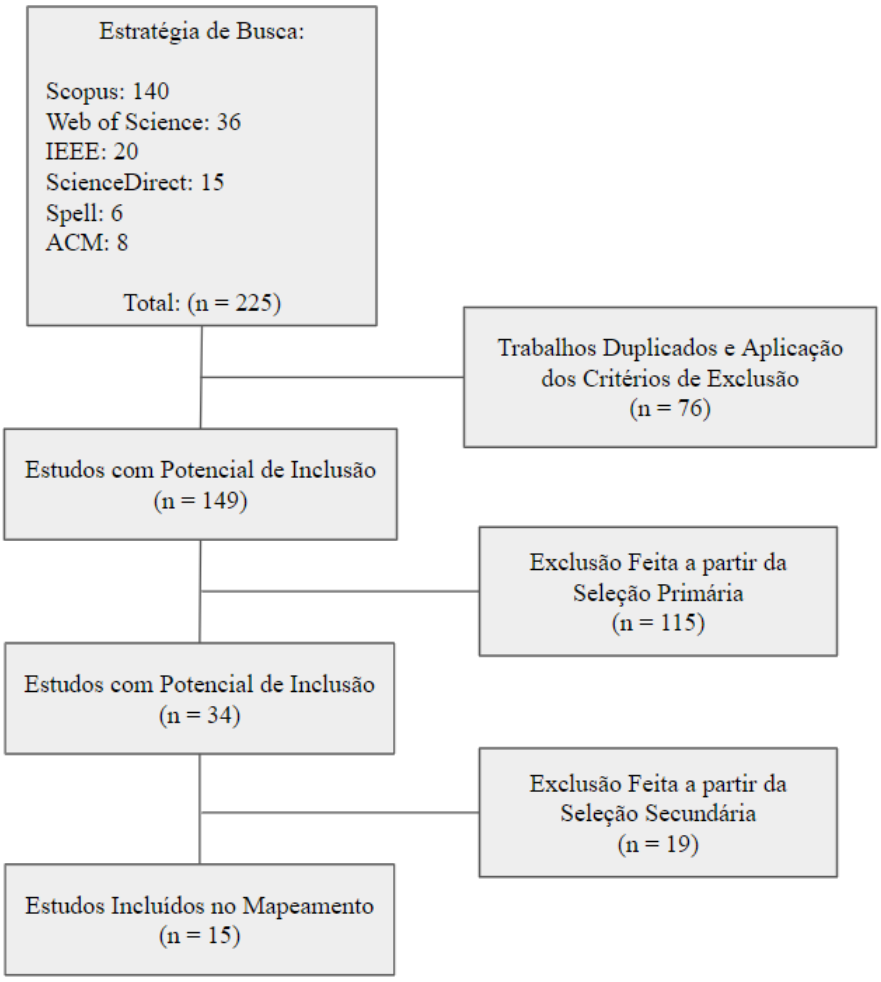


Figura 4 – Gráfico de Prisma com a extração dos Dados.

2.3.1 Quais são as abordagens utilizadas para mitigar casos de descrições textuais incongruentes em notas fiscais?

A Figura 5 apresenta a caracterização das principais abordagens de análise encontradas. Dentre as abordagens analisadas, o destaque vai para o aprendizado de máquina supervisionado, que esteve presente em 60% dos trabalhos.

2.3.2 Quais as principais técnicas utilizadas?

A Figura 6 apresenta a caracterização das principais técnicas utilizadas. Pode-se observar que as técnicas de Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN),

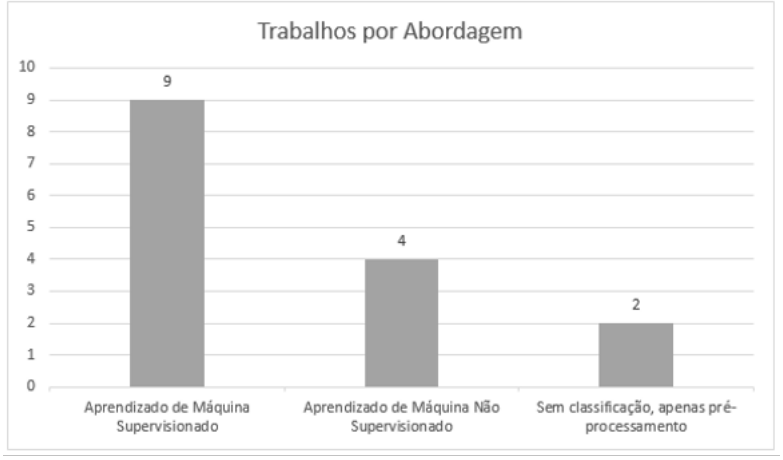


Figura 5 – Trabalhos Selecionados por Abordagem.

utilizando a arquitetura *Long Short Term Memory* (LSTM), foram as que mais se destacaram entre os trabalhos, tendo juntas, a presença em 40% dos artigos.

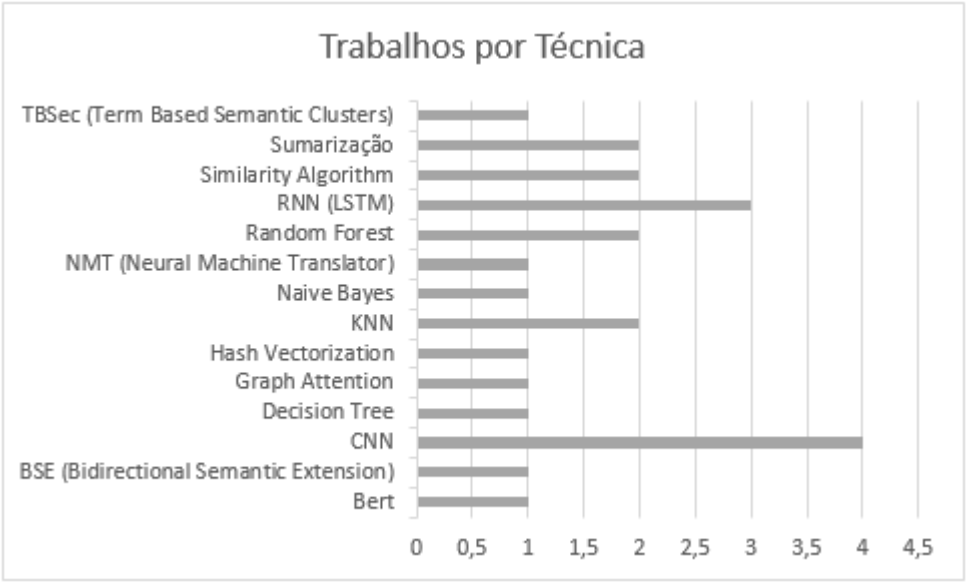


Figura 6 – Trabalhos Selecionados por Técnica.

2.3.3 Quais técnicas apresentam melhores eficácias?

Dentre as principais técnicas e as diversas formas de abordar o problema, podemos destacar a ferramenta apresentada em (DU et al., 2021), que combinou a técnica de Redes Neurais Recorrentes, utilizando a arquitetura LSTM, e um mecanismo de *Graph Attention* (GA), para alcançar uma acurácia de 93,10% em dados da indústria de veículos. Outra técnica que merece destaque é a *Bidirectional Semantic Extension* (BSE), apresentada no trabalho de Yue et al. (2020), que classificou textos curtos em notas fiscais chinesas, obtendo mais de 90% de acurácia.

2.3.4 Quais os meios de publicações mais populares?

A Figura 7 apresenta os trabalhos selecionados por tipo de publicação.

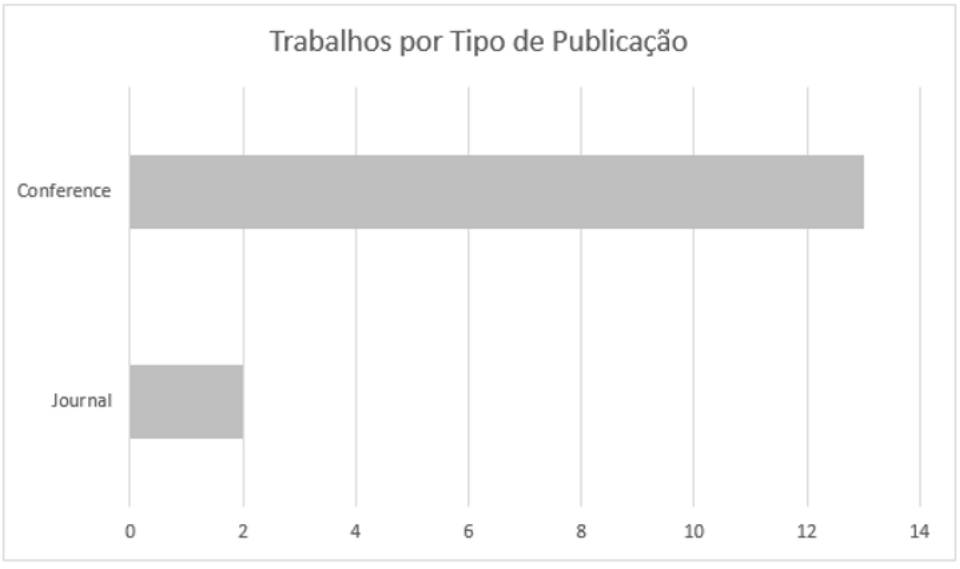


Figura 7 – Trabalhos Selecionados por Tipo de Publicação.

2.3.5 Em quais anos foram publicados mais artigos nesta área?

A Figura 8 mostra os trabalhos selecionados por ano de publicação. Pode-se observar que a maioria dos artigos foram publicados recentemente, no ano de 2021.

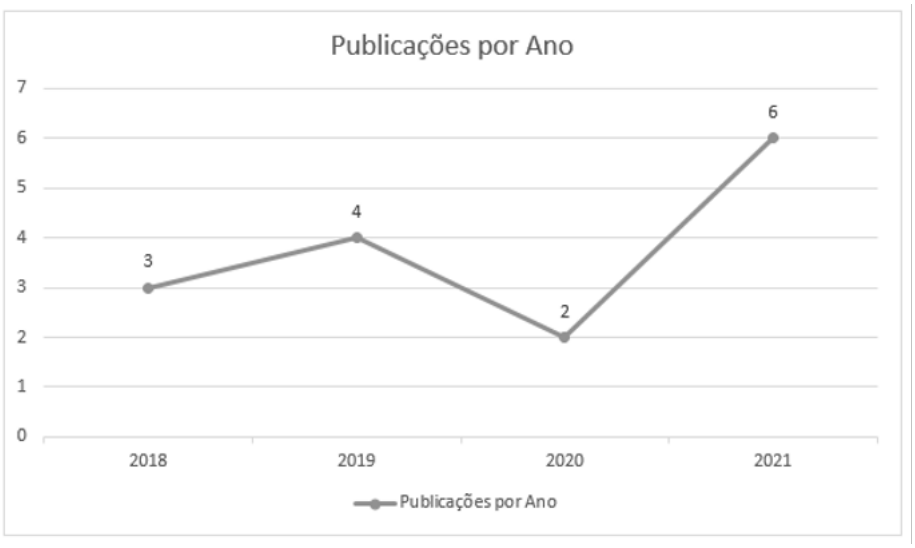


Figura 8 – Artigos Selecionados por Ano de Publicação.

2.3.6 Quais países têm mais publicações nesta área?

A Figura 9 apresenta os países que realizaram publicações sobre o tema abordado neste mapeamento. O país que mais se destacou foi a China com o maior número de publicações, seguido do Brasil e da Holanda.

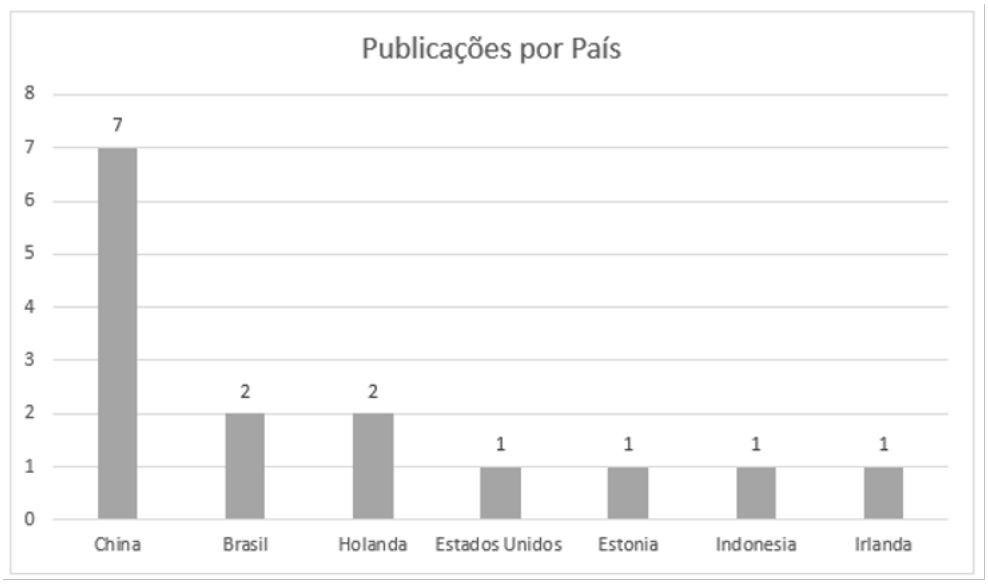


Figura 9 – Publicações por país.

2.4 Síntese Narrativa

Nesta seção, serão discutidos os principais aspectos encontrados e lições aprendidas nos artigos analisados. Diversas abordagens e técnicas inteligentes foram utilizadas para mitigação do problema de descrições textuais incongruentes em notas fiscais. O maior número de artigos publicados sobre o tema ocorreu em 2021, o que indica que a área de pesquisa ainda é jovem. Os resultados também mostram que as publicações relacionadas ao tema estão presentes em vários países, o que denota que a busca por uma solução em descrições incongruentes de notas fiscais apresenta-se como uma preocupação global.

A maioria dos trabalhos abordaram o problema como sendo de classificação, com a ideia de agrupar os produtos, ou classificar os produtos em seu código correto. Por exemplo, o trabalho de Du et al. (2021) propôs uma rede neural de classificação de código HS denominada HScodeNet, incorporando a informação hierárquica sequencial e espacial global dos textos, na qual um módulo de aprendizagem de sequência hierárquica é projetado para capturar a informação espacial de textos de descrição das mercadorias, utilizando *Graph Attention* (GA). A rede obteve 93,10% de acurácia em dados da indústria de veículos. Em outra dimensão, o trabalho de He et al. (2021) se concentra na identificação de categorias de commodities com base nas informações textuais nas declarações aduaneiras. Para isso, propôs um método de fusão de decisão simétrica baseado em Rede Neural Convolucional (CNN) e *Transformer* (BERT). Em (YU et al., 2019),

os autores propuseram um sistema de classificação baseado em um modelo de composição de Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN), com um mecanismo de atenção para recomendar as categorias correspondentes de comportamento de transação, que mapeia este comportamento na nota fiscal para o código da transação no documento de classificação e codificação tributária emitido pelo estado.

Paramartha et al. (2021) apresentaram um framework de aprendizado de máquina para classificar códigos HS, utilizando *K-fold Cross Validation*, com os classificadores *Naive Bayes, Decision Tree, Random Forest e K-Nearest Neighbor* (KNN). Inicialmente, eles replicaram o estudo de Altaheri & Shaalan (2020), utilizando os dados de importação da Indonésia, para posterior comparação com o framework gerado. O trabalho de Muir (2021) utilizou uma Rede Neural Convolucional (CNN), ao nível de caracter, com comprimento máximo limitado a 250, para classificar compras públicas em uma árvore taxonômica, por meio do uso de metadados. Os autores utilizaram quase 4 milhões de dados históricos de compras governamentais dos Estados Unidos, obtendo 48,50% de acurácia com a técnica *Top-Down*, quando o número de classes era 2.915.

Em (SPICHAKOVA; HAAV, 2020), os autores forneceram soluções automatizadas para o problema de classificação errada de mercadorias, por meio de uma abordagem híbrida que combina conhecimento derivado das descrições textuais e a taxonomia da nomenclatura dos códigos HS. Utilizando a técnica de similaridade do cosseno, verificaram se descrições textuais similares estão relacionadas a códigos HS similares. O trabalho de Yue et al. (2020) propôs um método de classificação de textos extremamente curtos em notas fiscais da China, nas quais a associação entre o nome e o rótulo da classe é fraca. O método foi baseado no *Bidirectional Semantic Extension*, e utilizou um mapa de conhecimento chinês para consultar sinônimos, ampliando assim o comprimento dos textos. A técnica alcançou mais de 90% de acurácia e se mostrou superior, quando comparada com outras técnicas tais como *LibShortText*, *NBSVM* e *TextGrocery*.

Alguns trabalhos apresentaram ideias diferentes para solução do problema. Por exemplo, Chen et al. (2021) utilizaram registros de declarações históricas para treinar um modelo não supervisionado que prevê uma pontuação a qual indica o grau de o código HS estar correto ou não, ao invés de classificar normalmente. Cada descrição é convertida para um vetor (*word2vec*) e a média dos vetores dentro do mesmo grupo de códigos HS é calculada e tomada como referência *(vector centroid)*. Já em (CHEN; BROMURI; EEKELEN, 2021), os autores solucionaram o problema por meio de uma Tradução Automática Neural (NMT), utilizando Redes Neurais Recorrentes (RNN), com uma arquitetura *encoder-decoder*, na qual o código HS representa o idioma alvo a ser produzido a partir dos recursos e descrição do item sendo enviado.

Como destaque no âmbito do Brasil, os trabalhos de Correa & Leal (2018) e Almeida et al. (2018) utilizaram técnicas de pré-processamento da mineração de textos para reclassificar as descrições dos produtos e serviços adquiridos pelo Governo Federal do Brasil.

Finalmente, diante da complexidade do problema, a qual varia de acordo com o tipo de produto a ser corretamente classificado, é importante destacar que a maioria dos trabalhos utilizaram uma combinação de técnicas para alcançar as melhores eficácias. Além disso, considerando novamente que os produtos possuem níveis de complexidade específicos para as suas descrições, avaliações por tipos de produtos poderão alcançar eficácias superiores às globais e soluções específicas para produtos mais importantes e rentáveis poderão ser viáveis.

2.5 Ameaças à Validade

Ameaças à validade podem limitar a habilidade de interpretar e/ou descrever resultados dos dados obtidos em um experimento (CHAPETTA, 2006). Portanto, algumas ameaças devem ser levadas em consideração.

Validade de Construção: A *string* de busca e as perguntas de pesquisa elaboradas podem não abranger toda a área de estudos relacionados à classificação de descrições textuais incongruentes, por meio do uso de inteligência artificial. Para mitigar tal ameaça, procurou-se elaborar uma *string* baseando-se em termos identificados e refinados com o auxílio dos artigos de controle norteados no modelo PICO.

Validade Interna: Para mitigar problemas de extração e caracterização dos dados, a análise de seleção e extração foi revisada por dois pesquisadores. Dúvidas foram resolvidas por meio de votação e da consulta a um terceiro especialista.

3

Prova de Conceito de um Classificador de OPMEs em Notas Fiscais

3.1 Materiais e métodos

Inicialmente, foi implementada a ferramenta OPMinEr. O algoritmo utilizado nesta primeira versão foi o *Naïve Bayes Multinomial*, com Frequência Inversa, o qual foi selecionado baseado nos resultados de um mapeamento sistemático sobre classificação de descrições associadas a códigos errados (GOMES; COLAÇO JÚNIOR, 2022), e nos resultados apresentados em (SANTOS et al., 2015), os quais evidenciaram, entre os algoritmos mais eficientes para mineração de textos, a boa eficácia desta abordagem.

De fato, a ferramenta realiza três classificações, em primeiro lugar é feita a predição do Tipo de OPME: Órtese, Prótese ou Material Especial. Neste caso, quando uma Nota não é uma OPME, esta é classificada como Outros. A próxima classificação indica à qual classe pertence o produto. Existem 83 classes diferentes: Stent; Cateter; Prótese Ortopédica; Cânula, etc. Por fim, é realizada a última classificação, a qual prediz o Procedimento do Sistema Único de Saúde (SUS) no qual o produto pode ter sido usado. Existem 319 procedimentos diferentes: Balão Destacável; Cateter Atrial / Peritoneal; Fio de Kirschner; Haste Intramedular Tíbio-Tarsica, etc.

Ato contínuo, foi realizada uma análise e discussão da ferramenta, utilizando o método de prova de conceito. Uma prova de conceito (PoC), do inglês *proof of concept*, é a evidência de que um produto ou um serviço potencial pode atender determinado objetivo, documentada a partir de testes, servindo para verificar e validar um projeto antes que este seja executado e disponibilizado para uso geral (CRUZ; COLAÇO JÚNIOR; GOIS, 2022). Na próxima seção, de forma autocontida, a prova de conceito e sua metodologia são detalhadas.

3.2 Prova de Conceito

3.2.1 Objetivo

Avaliar preliminarmente a ferramenta OPMinEr, na tarefa de identificação e classificação de notas fiscais de OPMEs.

3.2.2 Planejamento

3.2.2.1 Seleção de Participantes

Os dados utilizados para a realização da prova de conceito foram obtidos de notas fiscais disponibilizadas pelo governo do Rio Grande do Norte (RN), custodiadas pelo AudSUS, Laboratório de Inovação Tecnológica em Saúde (LAIS) e Ministério Público Federal. Ao todo, foram 465.726 notas fiscais, que correspondem ao período de janeiro de 2020 a maio de 2022. Pelo fato de existirem bem mais notas de produtos em geral, os quais não são OPME, foi selecionada uma amostra de 4.549 registros, mantendo uma proporção aproximada da quantidade de notas de OPMEs.

3.2.2.2 Dicionário (Modelo de Conhecimento)

O modelo de conhecimento utilizado para aprendizado de como as OPMEs são descritas foi composto com registros oriundos da Agência Nacional de Vigilância Sanitária (Anvisa), cadastro oficial de materiais do governo federal (Catmat), descrições avulsas (inseridas manualmente) escritas por auditores e descrições das próprias notas fiscais do Rio Grande do Norte (RN), rotuladas manualmente. A Similaridade do Cosseno foi utilizada para localizar OPMEs de forma semiautomática nos dados da Anvisa e do Catmat, em comparação com os procedimentos realizados pelo SUS. Descrições com Cosseno acima de 0,6, de um máximo de 1, foram conferidas manualmente e inseridas no dicionário.

Além disso, um dicionário de conflação foi criado, o qual uniformiza palavras. Por exemplo, quando a palavra “parafus.” é encontrada numa nota, esta é automaticamente transformada em “parafuso”.

3.2.2.3 Instrumentação

Os materiais e recursos utilizados foram:

- Ferramenta OPMinEr; PostgreSQL 12.12; Python 3.10.8; Django 4.1.7

3.2.3 Operação

3.2.3.1 Preparação

Em síntese, foi preparado o ambiente para a realização da prova de conceito, ou seja, o carregamento de todos os dados para o banco de dados.

3.2.3.2 Execução

O processo foi iniciado pelas tarefas de normalização e conflação das descrições, com intuito de padronização e mitigação de erros de classificação. A normalização, neste trabalho, é responsável por manter o texto com caracteres minúsculos e por remover, caso exista na descrição, os seguintes tokens ou caracteres especiais: ("\'[!@#$%^&*()[]{};:.<>?|‘~=_-); data; acentuação, mantendo a letra original; espaços em branco em excesso; pontuação e *stopwords* (preposição, artigos, etc). Ato contínuo, foi iniciado o processo de transformação dos dados em um modelo vetorial. Após a transformação dos dados em um modelo vetorial, foi realizado o treinamento e avaliação do modelo, utilizando uma técnica chamada *k-fold cross validation*. O processo de validação cruzada envolve a divisão do conjunto de dados disponível em várias partes (*folds*), geralmente em k partes iguais. Para cada iteração do processo de validação cruzada, uma das k partes é usada como conjunto de teste, enquanto as outras k-1 partes restantes são usadas como conjunto de treinamento. Neste trabalho, os dados do dicionário foram divididos em 10 conjuntos (10-*fold*), ou seja, o modelo foi treinado e testado dez vezes, perfazendo, para cada métrica avaliada, a média das 10 interações. Vale ressaltar que como a ferramenta realiza três tipos de classificação diferentes, foi necessário construir e treinar três modelos distintos. Ao término da execução, foram obtidas as seguintes métricas: acurácia, sensibilidade, medida-F1, precisão e tempo médio de execução. Os resultados desses dados coletados serão apresentados na próxima seção.

3.2.4 Resultados

Após a etapa de execução, foram obtidos o tempo médio e métricas para cada tipo de classificação. A Tabela 5 apresenta o tempo médio de execução de todo o processo de classificação, para cada modelo. O tempo médio de pré-processamento é o tempo levado para pré-processar as descrições e gerar o modelo vetorial, sendo o mesmo para todos os modelos, uma vez os modelos utilizam os mesmos dados de entrada. É possível perceber que, apesar do ótimo desempenho apresentado, quanto maior o número de classes alvo, maiores são os tempos de treinamento e classificação.

A Tabela 6 apresenta as métricas após a execução da classificação, para os três modelos. A ferramenta teve um ótimo desempenho em relação às métricas obtidas. A abordagem obteve uma acurácia de 99% em relação ao modelo para classificar o Tipo de OPME. Isso já assegura a recuperação de quase todas as notas de OPME, permitindo a localização completa de produtos

	Pré Processamento	**Treinamento**	**Classificação**
Tipo	1,71 s	5,09 s	0,007 s
Classe	1,71 s	7,62 s	0,014 s
Procedimento	1,71 s	12,00 s	0,044 s

Tabela 5 – Tempo médio de execução.

Fonte: Autores.

considerados mais importantes. Numa linha mais sofisticada de classificação, o modelo de associação a Procedimentos do SUS obteve uma boa acurácia, 92%, e uma precisão de 93%, mesmo considerando uma classificação para 319 Procedimentos distintos.

	Acurácia	**Sensibilidade**	**Medida-F1**	**Precisão**
Tipo	0,99	0,99	0,99	0,99
Classe	0,97	0,97	0,97	0,98
Tipo	0,92	0,92	0,92	0,93

Tabela 6 – Métricas do Processamento.

Fonte: Autores.

Por fim, vale ressaltar que existe a hipótese de melhorias dos resultados, considerando que os dicionários serão incrementados com notas de todo o Brasil.

4

Um Classificador Inteligente de OPMEs em Notas Fiscais

4.1 Base Conceitual

4.1.1 Controle da Atividade Pública

Dos tributos sob a responsabilidade da esfera estadual, o ICMS é a principal fonte de receita própria, contribuindo com significativos 19,7% da arrecadação total do país. Este imposto é não-cumulativo, permitindo a compensação do valor devido em cada transação relacionada à circulação de mercadorias ou à prestação de serviços com o montante cobrado em transações anteriores, conforme informações do Instituto Brasileiro de Planejamento e Tributação (IBPT) (IBPT, 2019).

Neste contexto, a Nota Fiscal Eletrônica é uma das principais provas de validade jurídica das operações comerciais realizadas pelo contribuinte. Sendo um programa de âmbito nacional, a Nota Fiscal Eletrônica (NF-e) foi desenvolvida pela Administração Tributária, a qual instituiu um modelo unificado de documento fiscal em meio eletrônico, substituindo o documento fiscal em papel. A NF-e tem como principal objetivo a modernização da Administração Tributária por meio da redução de custos e entraves burocráticos, facilitando o cumprimento das obrigações tributárias pelo contribuinte, além de fortalecer o controle e fiscalização pelos órgãos de controle (BRASIL, 2003).

A representação gráfica (física) da NF-e é denominada de Documento Auxiliar da Nota Fiscal Eletrônica (DANFE). O DANFE possui os mesmos campos definidos nos modelos anteriores à NF-e, este serve apenas como um instrumento auxiliar na consulta à NF-e, pois contém impressa a chave de acesso, a qual permite a validação do documento no site da Sefaz. O DANFE não é uma nota fiscal e nem a substitui.

A partir do DANFE é possível identificar os bens que foram faturados. Estes itens, produtos ou serviços possuem um código identificador que é utilizado pelo emissor da NF-e para

sua identificação em seu sistema de faturamento, uma descrição textual do item para leitura e fácil identificação ao que se refere o item e o código NCM - Nomenclatura Comum do Mercosul -, além das informações tributárias, quantidades e valores do bem (FEDERAL, 2023). Um NCM nada mais é que um número, cujo seu objetivo é facilitar a identificação da mercadoria para a fiscalização e tributação desta pelas autoridades competentes. Cada NCM possui um valor de alíquota atrelado, correspondente ao imposto que incidirá sobre o bem, além disso, cada item na nota deve ser vinculado obrigatoriamente ao código correspondente na tabela NCM.

Todavia, alguns contribuintes podem associar códigos referentes à tabela de NCMs que não correspondam, de fato, aos itens dos produtos descritos nas NF-e, de modo que pode haver a incidência de alíquotas erradas sobre os produtos fornecidos. Como apresentado por Carvalho et al. (2014), este fato impacta a arrecadação do Estado, uma vez que o imposto devido não é corretamente recolhido. Além disso, há a impossibilidade de autuação do contribuinte envolvido, por falta de meios que evidenciem a existência de irregularidades na emissão da NF-e. Finalmente, códigos incompatíveis ou incongruentes com as verdadeiras descrições dos produtos podem inviabilizar as investigações de órgãos de controle, que ficam impedidos de averiguar com precisão variáveis como preço médio, valores e quantidades exorbitantes de produtos adquiridos, muitas vezes contabilizados em categorias totalmente diferentes das suas.

4.1.2 Medida de Importância de uma Palavra (TF-IDF)

No momento do pré-processamento, o texto é quebrado em palavras, que poderão ser chamadas de *tokens*. Os *tokens* são quebrados, de modo geral, pelos espaços, e cada palavra que esteja entre espaços torna-se um *token*. Este processo de tokenização é utilizado para identificar palavras-chave que façam sentido e representem o documento (VIJAYARANI; JANANI, 2016).

A frequência do Termo, do inglês *Term-Frequency (TF)*, corresponde ao número de vezes que o termo aparece no documento. Os termos que são frequentemente mencionados em determinados documentos podem servir como discriminantes. Para um cálculo mais contextual, a medida estatística TF-IDF, apresentada por Salton, Fox e Wu (1983), leva em conta a importância de uma palavra no *corpus* (dicionário completo de palavras), estando estruturada ou não. Para cálculo do TF-IDF, é atribuído um valor para cada termo a partir da frequência do termo no próprio documento (TF já visto) e em todo o *corpus* (IDF), indicando sua importância.

A importância do termo é calculada pela equação de Frequência Inversa do Documento, *Inverse Document-Frequency (IDF)* (SALTON; BUCKLEY, 1988). O IDF é definido pela equação a seguir, a qual consiste no logaritmo da razão entre o número total de documentos, N_D, e a frequência de documentos que possuam o termo t, df_t. Quanto maior o IDF do termo, maior é a sua representatividade.

$$IDF_t = \log(\frac{N_D}{df_t})$$

O peso final do termo é atribuído pela equação do TF-IDF, na qual o peso é associado à proporção da frequência do termo no documento (TF) e à proporção inversa do número de no qual o termo aparece pelo menos uma vez (IDF). O TF-IDF é representado pela equação a seguir:

$$TF - IDF = TF \times IDF$$

Caso a palavra apareça em todos os documentos, seguindo a fórmula, a IDF será o Log de 1, ou seja, será zero, pois o Log de 1 em qualquer base é zero. Como o IDF será zero, quando multiplicado por qualquer valor do TF, o resultado também será zero. Em outras palavras, quanto mais a palavra aparecer no *corpus*, menor será o seu peso, podendo ser zero, caso esteja presente em todos os documentos (pouco poder de discriminação).

4.1.3 Métricas de Qualidade

Para avaliação do classificador, foram utilizadas as métricas: acurácia, sensibilidade, precisão e medida-F1 (ZHU; ZENG; WANG, 2010). Essas medidas são mensuradas a partir das seguintes frequências:

- *True Positive (TP)*: Total de instâncias de produtos das notas, presentes na base anotada (base de treinamento do algoritmo), que foram classificadas corretamente;
- *True Negative (TN)*: Total de instâncias de produtos, que não são OPMEs, e foram classificadas corretamente;
- *False Positive (FP)*: Total de instâncias de outros produtos que foram classificados incorretamente como pertencentes à OPME;
- *False Negative (FN)*: Total de instâncias de produtos de OPME que não foram classificados corretamente;

4.1.3.1 Acurácia

A acurácia (*accuracy*) representa o percentual de instâncias (notas fiscais) que foram classificadas de forma correta, sendo definida por:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

4.1.3.2 Sensibilidade

A sensibilidade ou revocação (*recall*), também conhecida como a taxa de verdadeiros positivos, sensibilidade ou cobertura real da amostragem positiva, é o percentual de instâncias positivas que foram classificadas corretamente:

$$recall = \frac{TP}{TP+FN}$$

4.1.3.3 Precisão

A precisão (*precision*) é a razão entre as instâncias classificadas como "verdadeiro positivo" e todas as instâncias classificadas como positivas:

$$precision = \frac{TP}{TP+FP}$$

4.1.3.4 Medida-F1

A medida-F1 (*F1-Score*) é a métrica que combina dois indicadores de desempenho, sendo a expressão da média harmônica da precisão e da sensibilidade:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

4.2 Metodologia

O mapeamento sistemático permitiu a identificação dos algoritmos mais utilizados, mais eficazes e mais rápidos, viabilizando uma seleção dos que se destacaram para estas duas últimas características, uma vez que o ambiente da ferramenta aqui proposta possui características Big Data.

Do ponto de vista da classificação metodológica principal, este artigo é um estudo experimental, o qual segue os passos apresentados em (COLAÇO JÚNIOR et al., 2022), para avaliação dos resultados das classificações obtidas a partir do classificador de OPMEs, o **OPM**in**E**r. Esta avaliação, detalhada a seguir, utilizou as métricas de qualidade apresentadas na subseção 4.1.3, as quais foram implementadas a partir das bibliotecas mencionadas na subseção 4.4.2.6.

Para a concepção do modelo de conhecimento, foi construído um dicionário de dados, com informações de diversas fontes (vide seção 4.3), que serviu para treinamento dos algoritmos avaliados. Ato contínuo, foi executada a etapa de pré-processamento, a qual quebra as descrições em *tokens*. O modelo vetorial TF-IDF foi utilizado para calcular a frequência e a importância de um *token* no dicionário. Após o desenvolvimento do **OPM**in**E**r, o qual agrega o dicionário criado e disponibiliza todos os algoritmos selecionados no mapeamento, foi executado um experimento para avaliar estes algoritmos, do tipo *in vitro*, uma vez que os objetos foram retirados da base original para serem manipulados dentro de um ambiente controlado. Este método possibilitou analisar e avaliar os resultados utilizando análise estatística, bem como permitirá uma replicação mais fiel dos procedimentos apresentados neste trabalho.

A replicação de experimentos é uma característica importante para qualquer área da ciência. Assim, na área de software, devem haver métodos que permitam a replicação e a avaliação dos trabalhos, a fim de evitar que novos métodos, técnicas, linguagens e ferramentas

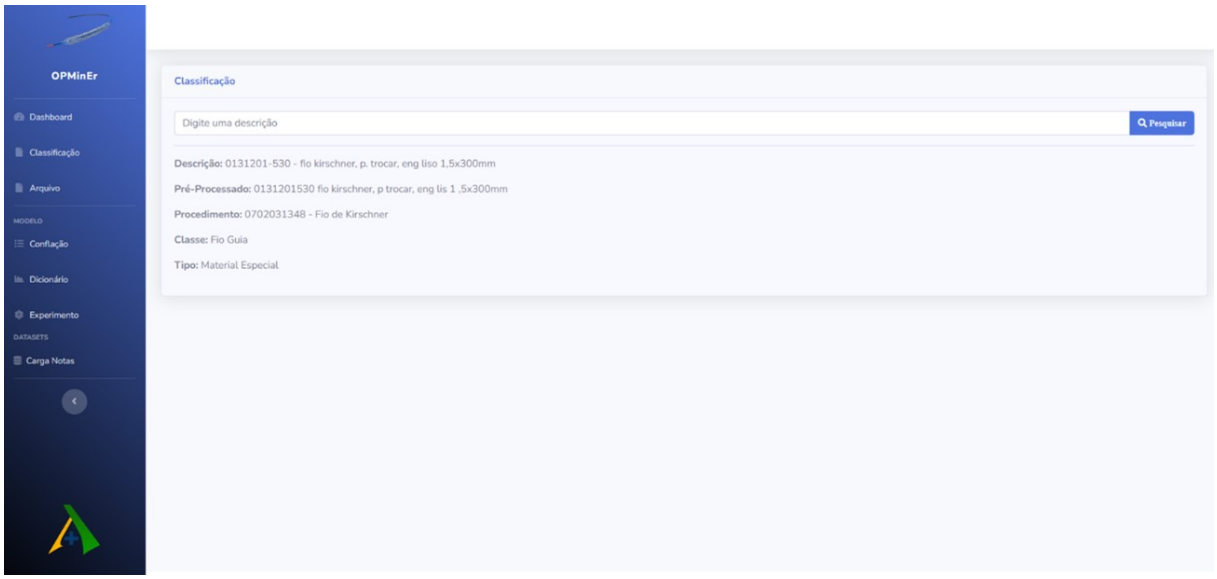
sejam sugeridos, publicados ou apresentados para venda, sem experimentação e validação (TRAVASSOS, 2002).

Por fim, a definição e o planejamento do experimento estão descritos detalhadamente, com suas **metodologias autocontidas**, na Seção 4.4. Em resumo, o experimento está dividido em quatro etapas principais: o planejamento; a operação de limpeza dos dados, coleta e geração da base experimental; a comparação de métodos; e a análise dos resultados.

4.3 OPMinEr

O **OPM**in**E**r é um sistema que executa três tipos de classificações relacionadas a OPMEs. Essas classificações são realizadas em sequência. Primeiro, o sistema faz a predição do tipo de OPME (Órtese, Prótese ou Material Especial) com base na descrição do produto. Se a descrição não se encaixar em nenhuma dessas categorias, o produto é classificado como "Outros". Em seguida, o sistema classifica o produto em uma das 83 classes diferentes, como Stent, Cateter, Prótese Ortopédica, Cânula, etc. Por fim, o OPMinEr prevê o procedimento específico do Sistema Único de Saúde (SUS) no qual o produto pode ter sido utilizado. Existem 319 procedimentos diferentes nessa categoria, tais como procedimentos para colocação de Balão Destacável, Cateter Atrial / Peritoneal, Fio de Kirschner, Haste Intramedular Tíbio-Tarsica, entre outros. Essas classificações são realizadas após o pré-processamento da descrição do produto, e o OPMinEr utiliza modelos previamente treinados para executar essas tarefas. A Figura 10 mostra um exemplo do processo de classificação. A operação completa está descrita na seção 4.5.

Figura 10 – Tela de Classificação.



Fonte: Autores.

O modelo de conhecimento utilizado para aprender como as OPMEs (Órteses, Próteses e

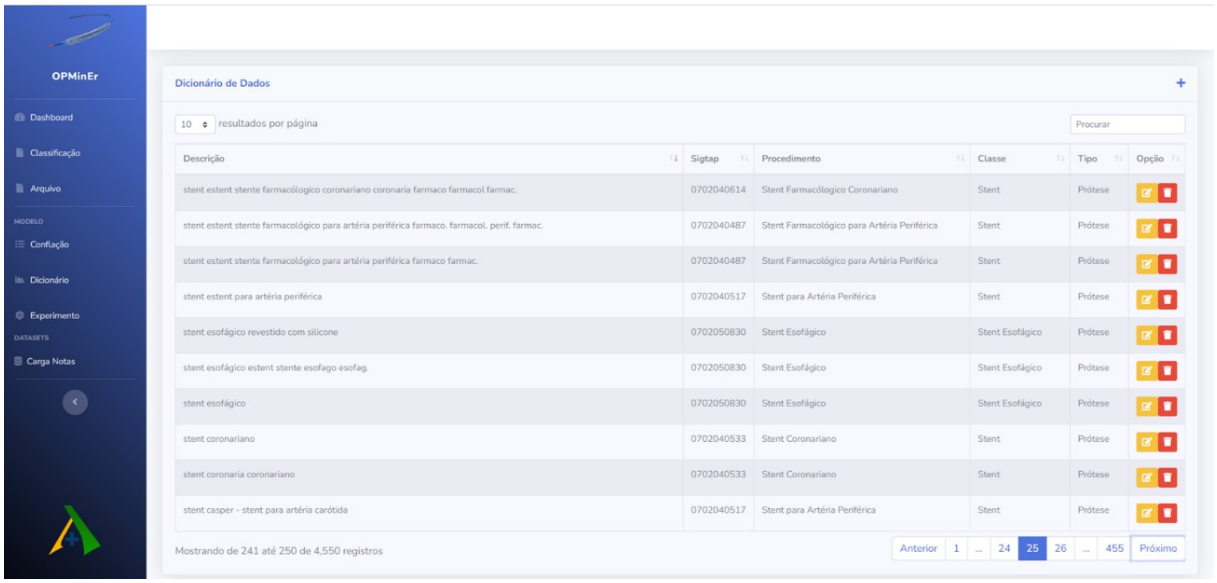
Materiais Especiais) são descritas foi construído com uma variedade de fontes de dados. Essas fontes incluíram registros da Agência Nacional de Vigilância Sanitária (Anvisa), o cadastro oficial de materiais do governo federal (Catmat), descrições adicionais inseridas manualmente por auditores e informações provenientes das próprias notas fiscais do estado do Rio Grande do Norte (RN), que foram rotuladas manualmente.

Um *script* em linguagem *Python* foi desenvolvido para automatizar o download dos arquivos .csv do Catmat e inseri-los em uma tabela de banco de dados, facilitando a criação de um repositório de informações atualizado e acessível.

A Similaridade do Cosseno, uma técnica de processamento de linguagem natural, foi empregada para identificar, de forma semiautomática, produtos presentes nas bases de dados da Anvisa e do Catmat que se encaixam na categoria de OPMEs. O *script* lê os dados dos procedimentos do SUS referentes a OPMEs, e calcula a similaridade do cosseno para as descrições dos produtos da Anvisa e Catmat. Descrições com uma similaridade do cosseno menor ou igual a 0,6 foram inicialmente descartadas como parte de um filtro preliminar. Posteriormente, as descrições eleitas passaram por uma revisão manual e foram inseridas no dicionário de dados. A Figura 11 do artigo apresenta algumas dessas descrições contidas no dicionário, ilustrando a diversidade e a complexidade das descrições de OPMEs.

Além disso, um dicionário de conflação e padronização de palavras foi criado como parte do sistema. Este dicionário não apenas uniformiza palavras, mas também serve como uma ferramenta eficaz de sinônimos. Por exemplo, sempre que o sistema encontra a palavra "parafus.", ele realiza uma correção automática, transformando-a instantaneamente em "parafuso".

Figura 11 – Tela de Dicionário de Dados.

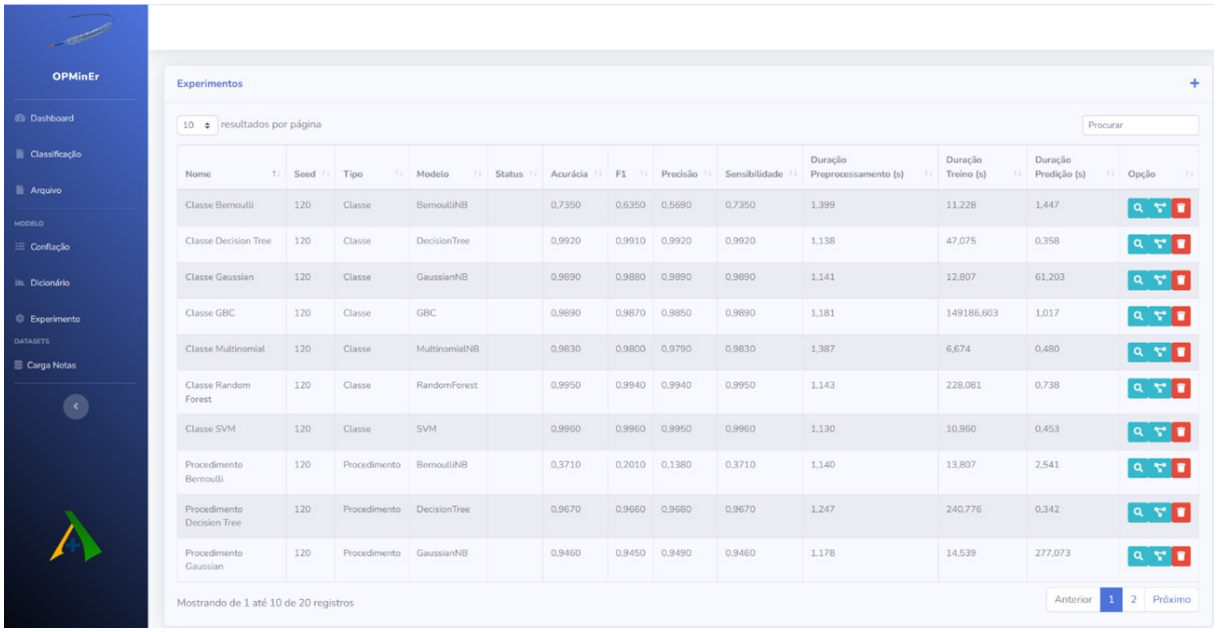


Fonte: Autores.

A Figura 12 apresenta a tela dos experimentos executados pelo OPMinEr. É relevante

observar que diferentes classificações nas notas fiscais podem envolver algoritmos distintos. Além disso, a ferramenta está apta a ser reavaliada conforme o dicionário é incrementado ao longo do tempo.

Figura 12 – Tela de Experimentos.



Fonte: Autores.

4.4 Definição e Planejamento do Experimento

4.4.1 Objetivo

Analisar um classificador de OPMEs, o OPMinEr, o qual se baseia em processamento de linguagem natural, **com a finalidade de** avaliar, **com relação à** acurácia, precisão, sensibilidade, medida F-1, tempo médio de treinamento e execução, **do ponto de vista de** auditores, profissionais de saúde e cientistas de dados, **no contexto de** descrições de notas fiscais, analisadas pelos Ministérios Público e da Saúde, bem como pela Auditoria do Sistema Único de Saúde Brasileiro (AudSUS).

4.4.2 Planejamento

4.4.2.1 Seleção de Contexto

O experimento foi realizado "in vitro", análise em um ambiente controlado, no qual foi utilizado o OPMinEr para classificar OPMEs em notas fiscais. Foram obtidos dados de notas fiscais custodiadas pelo AudSUS, Laboratório de Inovação Tecnológica em Saúde (LAIS) e

Ministério Público Federal. Ao todo, foram 465.726 notas fiscais, que correspondem ao período de janeiro de 2020 a maio de 2022.

4.4.2.2 Questões de Pesquisa

No contexto de classificação de OPMEs em notas fiscais, dos algoritmos selecionados, duas questões envolvem a formulação das hipóteses, são elas:

- Q1: Entre os algoritmos selecionados, qual o melhor em termos de eficácia?
- Q2: Entre os algoritmos selecionados, qual o melhor em termos de eficiência?

Formulação de Hipóteses: Para responder a questão Q1, as seguintes hipóteses foram elaboradas:

- H_0: Os algoritmos (1,2..n) têm a mesma eficácia.
- H_1: Os algoritmos (1,2..n) têm eficácias diferentes.

Para responder a questão Q2, as seguintes hipóteses foram elaboradas:

- H_0: Os algoritmos (1,2..n) têm a mesma eficiência.
- H_1: Os algoritmos (1,2..n) têm eficiências diferentes.

4.4.2.3 Variáveis Independentes

As variáveis independentes consideradas para esta experimentação foram: dicionário de dados construído; base de notas fiscais eletrônicas com as descrições alvo a serem classificadas; o classificador OPMinEr, os algoritmos utilizados para tarefa de classificação: *Naïve Bayes Multinomial*, *Naïve Bayes Gaussian* e *Naïve Bayes Bernoulli*, *Linear Support Vector*, *Decision Tree*, *Random Forest* e *Gradient Boosting*.

4.4.2.4 Variáveis Dependentes

As predições realizadas, das quais as métricas podem ser derivadas: Acurácia (ACU), Precisão (PRE), Sensibilidade (SEN), Medida-F1 (F1). Mais ainda, o tempo médio de treinamento (TMT) e o tempo médio de classificação (TMC).

4.4.2.5 Seleção de Participantes e Objetos

Pelo fato de existirem bem mais notas de produtos em geral, os quais não são OPME, foi selecionada uma amostra de 4.718 registros, mantendo uma proporção aproximada da quantidade de notas de OPMEs. A escolha dos algoritmos de classificação foi baseada num mapeamento sistemático realizado por Gomes & Colaço Júnior (2022), os quais elencaram algoritmos de classificação mais rápidos para resolução do problema geral de descrições incongruentes.

4.4.2.6 Instrumentação

Os materiais e recursos utilizados foram:

- Ferramenta OPMinEr;
- *PostgreSQL* 12.12 (POSTGRESQL, 2021);
- *Python* 3.10.8 (PYTHON, 2021);
- *Django* 4.1.7 (DJANGO, 2023);
- *Mlflow* 2.3 (MLFLOW, 2023);
- *Power BI* 2.117.984.0 64-Bit (POWERBI, 2023);
- *Scikit-learn* (PEDREGOSA et al., 2012);
- Computador com Intel(R) Core(TM) i7-8700T CPU @ 2.40GHz e 32GB RAM;

4.5 Operação do Experimento

4.5.1 Preparação

Em síntese, foi preparado o ambiente para a realização do experimento controlado, ou seja, download e instalação dos itens descritos na seção 4.4.2.6, e carregamento de todos os dados para o banco de dados. Os dados de notas fiscais foram disponibilizados por meio de um dump de banco de dados, e os dados da ANVISA e CATMAT foram disponibilizados por meio de arquivos em formato .csv.

4.5.2 Execução

O processo foi iniciado pelas tarefas de normalização e conflação das descrições, com intuito de padronização e mitigação de erros de classificação. A normalização, neste trabalho, é responsável por manter o texto com caracteres minúsculos e por remover, caso exista na descrição, os seguintes *tokens* ou caracteres especiais: ("\'[!@#$%^&*()[];:.<>?|‘~=_-); data; acentuação,

mantendo a letra original; espaços em branco em excesso; pontuação e *stopwords* (preposição, artigos, etc). Ato contínuo, foi iniciado o processo de transformação dos dados em um modelo vetorial.

Após a transformação dos dados em um modelo vetorial, foi realizado o treinamento e avaliação do modelo, utilizando uma técnica chamada *k-fold cross validation*. O processo de validação cruzada envolve a divisão do conjunto de dados disponível em várias partes (*folds*), geralmente em k partes iguais. Para cada iteração do processo de validação cruzada, uma das k partes é utilizada como conjunto de teste, enquanto as outras k-1 partes restantes são usadas como conjunto de treinamento. Neste trabalho, os dados do dicionário foram divididos em 10 conjuntos (10-*fold*), ou seja, o modelo foi treinado e testado dez vezes, perfazendo, para cada métrica avaliada, a média das 10 interações. Vale ressaltar que como a ferramenta realiza três tipos de classificação diferentes, foi necessário construir e treinar três modelos distintos. Ao término da execução, foram obtidas as seguintes métricas: acurácia, sensibilidade, medida-F1, precisão, tempo médio de treinamento e tempo médio de classificação. Para melhor gerenciamento dos modelos de teste e produção, os modelos e métricas foram gerenciados pelo *MLflow*, uma ferramenta de código aberto para gerenciar o ciclo de vida de modelos de aprendizado de máquina. Os resultados desses dados coletados serão apresentados na próxima seção.

4.5.3 Validação dos Dados

Para análise, interpretação e validação, foram utilizados 05 (cinco) tipos de testes estatísticos: *Shapiro-Wilk, Kruskal-Wallis, Post-Hoc Dunn, Anova e Tukey HSD*. O teste de *Shapiro-Wilk* (SHAPIRO; WILK, 1965) foi utilizado para testar a normalidade dos dados, já os testes de *Kruskal-Wallis* (KRUSKAL; WALLIS, 1952) e *Post-Hoc Dunn* (DUNN, 1961) foram utilizados para a comparação das medianas da Acurácia e Medida-F1, nos casos em que o teste de normalidade indicou uma distribuição de dados não-normal. Para os dados normais e com homocedasticidade, o teste *Anova* verifica se a diferença entre as médias de dois ou mais grupos é significativa. Ato contínuo, o teste de *Tukey* é usado como complemento para testar todo e qualquer contraste entre 2 médias dos tratamentos (FIELD, 2009).

4.6 Resultados

Para responder às questões de pesquisas elencadas na seção 4.4.2.2, ocorreu a etapa de execução e foram obtidos os resultados das classificações, assim como seu tempo médio, para cada uma das três classificações: Tipo, Classe e Procedimento. Na Tabela 7, são apresentadas as médias das métricas para a classificação de Tipo.

Como é perceptível, os algoritmos obtiveram médias de acurácia similares, com maior destaque para o Linear Support Vector. Em relação ao tempo de treinamento, o algoritmo *Naïve Bayes Multinomial* foi o mais rápido e em relação ao tempo de classificação, os algoritmos *Linear*

Algoritmos	**ACU**	**SEN**	**PRE**	**F1**	**TMT**	**TMC**
Linear Support Vector	**0,998**	**0,998**	**0,998**	**0,998**	7,058 s	**0,273 s**
Random Forest	0,996	0,996	0,996	0,996	221,193 s	**0,273 s**
Decision Tree	0,995	0,995	0,995	0,995	64,196 s	0,318 s
Gradient Boosting	0,995	0,995	0,995	0,995	17231,072 s	0,446 s
Naïve Bayes Bernoulli	0,99	0,99	0,991	0,99	8,468 s	0,834 s
Naïve Bayes Multinomial	0,99	0,99	0,991	0,99	**4,375 s**	0,277 s
Naïve Bayes Gaussian	0,99	0,99	0,991	0,99	13,146 s	3,573 s

Tabela 7 – Comparativo das médias das métricas para classificação Tipo.

Fonte: Autores.

Support Vector e *Random Forest* obtiveram os melhores resultados. Apesar da boa acurácia dos algoritmos para identificar se a descrição de uma nota fiscal é órtese, prótese ou material especial, não é possível fazer afirmações sem evidências estatísticas suficientemente conclusivas. Neste sentido, definiu-se um nível de significância (α) de 0,05 para todo o experimento. Ao aplicar o teste de *Shapiro-Wilk*, para análise da normalidade da distribuição dos dados, foram obtidos os *p-values* apresentados na Tabela 8.

Algoritmos	**ACU**	**F1**
Naïve Bayes Gaussian	0,899	0,899
Gradient Boosting	0,507	0,507
Random Forest	0,129	0,156
Decision Tree	0,093	0,093
Linear Support Vector	0,077	0,077
Naïve Bayes Bernoulli	0,073	0,073
Naïve Bayes Multinomial	0,014	0,014

Tabela 8 – Resultado do Teste de *Shapiro-Wilk*, para análise da normalidade dos dados de Tipo.

Fonte: Autores.

Uma vez que o teste de normalidade indicou distribuições não-normais para o caso do algoritmo *Naïve Bayes Multinomial* (*p-value* < 0,05), foi aplicado o teste não paramétrico de *Kruskal-Wallis*, com o objetivo de verificar se há pelo menos uma diferença entre as medianas das métricas. O teste foi aplicado para a Acurácia e para Medida-F1, salientando que esta última harmoniza as métricas de precisão e sensibilidade. Com a aplicação do teste de Kruskal-Wallis, verificou-se um *p-value* de 0.000143, fortemente menor que o nível de significância adotado,

tanto para métrica de Acurácia quanto para a Medida-F1. Desta forma, foi possível confirmar a evidência das diferenças entre as medianas, ou seja, as hipóteses H_0, de que os métodos possuem a mesma Medida-F1 e mesma Acurácia, foram rejeitadas. Em outras palavras, a diferença entre as medianas das classificações de alguns grupos é grande o suficiente para ser estatisticamente significativa.

Assim sendo, foi evidenciado que ao menos um método se diferencia dos demais, porém, não é possível afirmar qual é o mais discrepante. Para isso, foi utilizado o teste de *Kruskal-Wallis* dois a dois, equivalente ao teste de *Mann-Whitney U* com aproximação normal. A Tabela 9 apresenta os resultados deste teste, evidenciando que, após uma análise, e aplicação do teste *Post-Hoc Dunn*, aplicando a correção de *Bonferroni* ($\alpha = 0{,}0024$), para Acurácia e para Medida-F1, os seguintes pares foram significativamente diferentes (*p-value* < 0,0024): *Naïve Bayes Bernoulli - Linear Support Vector; Gaussian - Linear Support Vector; Multinomial - Linear Support Vector*.

A Tabela 10 apresenta as médias das métricas para a classificação de Classe.

Os algoritmos mantiveram acurácias similares, tendo apenas o *Naïve Bayes Bernoulli* se desviando dos demais. Assim como na classificação de Tipo, o algoritmo *Linear Support Vector* teve uma acurácia superior aos demais. Assim como na classificação de Tipo, o algoritmo *Naïve Bayes Multinomial* se destacou quanto a eficiência do tempo de treinamento, e o *Decision Tree* quanto ao tempo de classificação. Ao analisar a normalidade da distribuição dos dados, percebemos que para a classificação de Classe, os dados assumem uma distribuição não-normal para a Acurácia, e uma distribuição normal para a Medida-F1. A Tabela 11 mostra o teste de *Shapiro-Wilk* para Classe.

Com isso, foi possível utilizar o teste de *Anova* para a Medida-F1, com objetivo de verificar se a diferença entre as medianas das métricas de dois ou mais grupos é significativa. Após a aplicação do teste, verificou-se um *p-value* de 0,000. Desta forma, a hipótese H_0, de que os métodos possuem a mesma Medida-F1 foi rejeitada. O teste mostra que algumas das medianas dos grupos não são iguais, porém não diz quais são as diferenças. Para isso, foi aplicado o teste de *Tukey HSD* dois a dois e os seguintes pares foram significativamente diferentes: *Naïve Bayes Bernoulli - Decision Tree; Naïve Bayes Bernoulli - Naïve Bayes Gaussian; Naïve Bayes Bernoulli - Gradient Boosting; Bernoulli - Naïve Bayes Multinomial; Naïve Bayes Bernoulli - Random Forest; Bernoulli - Linear Support Vector; Decision Tree - Naïve Bayes Multinomial; Naïve Bayes Multinomial - Random Forest; Naïve Bayes Multinomial - Linear Support Vector*.

Visto que o teste de normalidade indicou distribuições não-normais para o caso do algoritmo *Naïve Bayes Bernoulli* para a medida de Acurácia, foi utilizado o teste de *Kruskal-Wallis* para verificar se a diferença entre as medianas das métricas é significativa. O *p-value* encontrado após a aplicação do teste foi 0,000, e portanto a hipótese H_0, de que os métodos possuem a mesma Acurácia foi rejeitada. Posto isto, o teste de *Kruskal-Wallis* dois a dois, aplicando a correção de *Bonferroni* ($\alpha = 0{,}0024$), foi utilizado para identificar os pares significativamente diferentes: *Naïve Bayes Bernoulli - Decision Tree; Naïve Bayes Bernoulli - Gradient Boosting;*

Comparação Dois a Dois		
Algoritmos	**ACU**	**F1**
Naïve Bayes Bernoulli - Decision Tree	0,055	0,055
Naïve Bayes Bernoulli - Naïve Bayes Gaussian	0,093	0,093
Naïve Bayes Bernoulli - Gradient Boosting	0,027	0,027
Naïve Bayes Bernoulli - Naïve Bayes Multinomial	0,570	0,570
Naïve Bayes Bernoulli - Random Forest	0,004	0,004
Naïve Bayes Bernoulli - Linear Support Vector	0,000	0,000
Decision Tree - Naïve Bayes Gaussian	0,044	0,044
Decision Tree - Gradient Boosting	0,768	0,768
Decision Tree - Naïve Bayes Multinomial	0,176	0,176
Decision Tree - Random Forest	0,332	0,332
Decision Tree - Linear Support Vector	0,060	0,060
Naïve Bayes Gaussian - Gradient Boosting	0,021	0,021
Naïve Bayes Gaussian - Naïve Bayes Multinomial	0,507	0,507
Naïve Bayes Gaussian - Random Forest	0,003	0,003
Naïve Bayes Gaussian - Linear Support Vector	0,000	0,000
Gradient Boosting - Naïve Bayes Multinomial	0,099	0,099
Gradient Boosting - Random Forest	0,500	0,500
Gradient Boosting - Linear Support Vector	0,112	0,112
Naïve Bayes Multinomial - Random Forest	0,020	0,020
Naïve Bayes Multinomial - Linear Support Vector	0,001	0,001
Random Forest - Linear Support Vector	0,361	0,361

Tabela 9 – Teste de *Kruskal-Wallis*, dois a dois.

Fonte: Autores.

Naïve Bayes Bernoulli - Random Forest; Naïve Bayes Bernoulli - Linear Support Vector; Naïve Bayes Multinomial - Random Forest.

A Tabela 12 apresenta os resultados do teste.

A Tabela 13 apresenta as médias das métricas para a classificação de Procedimento.

O gráfico da Figura 13 apresenta a disposição das medidas Acurácia e Medida-F1 em relação aos algoritmos.

Algoritmos	**ACU**	**SEN**	**PRE**	**F1**	**TMT**	**TMC**
Linear Support Vector	**0,996**	**0,996**	**0,995**	**0,996**	10,960 s	0,453 s
Random Forest	0,995	0,995	0,994	0,994	228,081 s	0,738 s
Decision Tree	0,992	0,992	0,992	0,991	47,075 s	**0,358 s**
Naïve Bayes Gaussian	0,989	0,989	0,989	0,988	12,807 s	61,203 s
Gradient Boosting	0,989	0,989	0,985	0,987	149186,603 s	1,017 s
Naïve Bayes Multinomial	0,983	0,983	0,979	0,980	**6,674 s**	0,480 s
Naïve Bayes Bernoulli	0,735	0,735	0,569	0,635	11,228 s	1,447 s

Tabela 10 – Comparativo das médias das métricas para classificação de Classes.

Fonte: Autores.

Algoritmos	**ACU**	**F1**
Gradient Boosting	0,906	0,473
Random Forest	0,274	0,578
Linear Support Vector	0,255	0,117
Naïve Bayes Gaussian	0,156	0,117
Decision Tree	0,093	0,085
Naïve Bayes Multinomial	0,058	0,116
Naïve Bayes Bernoulli	0,042	0,103

Tabela 11 – Resultado do Teste de *Shapiro-Wilk*, para análise da normalidade dos dados de Classe.

Fonte: Autores.

Para a classificação de Procedimento, o algoritmo *Linear Support Vector* teve uma acurácia acima de 98%. Já em relação a eficiência, temos o *Naïve Bayes Multinomial* como o mais eficiente para o tempo médio de treinamento e o *Decision Tree* para o tempo médio de classificação. A Tabela 14 mostra o teste de *Shapiro-Wilk* para análise da normalidade da distribuição dos dados da classificação de Procedimento.

Ao analisar a normalidade da distribuição dos dados, percebemos que para a classificação de Procedimento, os dados assumem uma distribuição normal tanto para medida de Acurácia, quanto para Medida-F1, e com isso, foi utilizado o teste de *Anova* para verificar a diferença entre as medianas das métricas de dois ou mais grupos é significativa. Após a aplicação do teste, verificou-se um *p-value* de 0,000. Desta forma, a hipótese H_0, de que os métodos possuem a mesma Acurácia e Medida-F1 foram rejeitadas. O teste de *Tukey HSD* dois a dois foi utilizado

Comparação Dois a Dois		
Algoritmos	**ACU**	**F1**
Naïve Bayes Bernoulli - Decision Tree	0,000	0,000
Naïve Bayes Bernoulli - Naïve Bayes Gaussian	0,003	0,000
Naïve Bayes Bernoulli - Gradient Boosting	0,001	0,000
Naïve Bayes Bernoulli - Naïve Bayes Multinomial	0,101	0,000
Naïve Bayes Bernoulli - Random Forest	0,000	0,000
Naïve Bayes Bernoulli - Linear Support Vector	0,000	0,000
Decision Tree - Naïve Bayes Gaussian	0,260	0,958
Decision Tree - Gradient Boosting	0,377	0,775
Decision Tree - Naïve Bayes Multinomial	0,013	0,025
Decision Tree - Random Forest	0,220	0,978
Decision Tree - Linear Support Vector	0,072	0,851
Naïve Bayes Gaussian - Gradient Boosting	0,808	0,999
Naïve Bayes Gaussian - Naïve Bayes Multinomial	0,179	0,247
Naïve Bayes Gaussian - Random Forest	0,018	0,535
Naïve Bayes Gaussian - Linear Support Vector	0,003	0,275
Gradient Boosting - Naïve Bayes Multinomial	0,113	0,515
Gradient Boosting - Random Forest	0,035	0,260
Gradient Boosting - Linear Support Vector	0,007	0,105
Naïve Bayes Multinomial - Random Forest	0,000	0,001
Naïve Bayes Multinomial - Linear Support Vector	0,000	0,000
Random Forest - Linear Support Vector	0,569	0,999

Tabela 12 – Teste de *Kruskal-Wallis* (ACU) e *Tukey HSD* (F1), dois a dois.

Fonte: Autores.

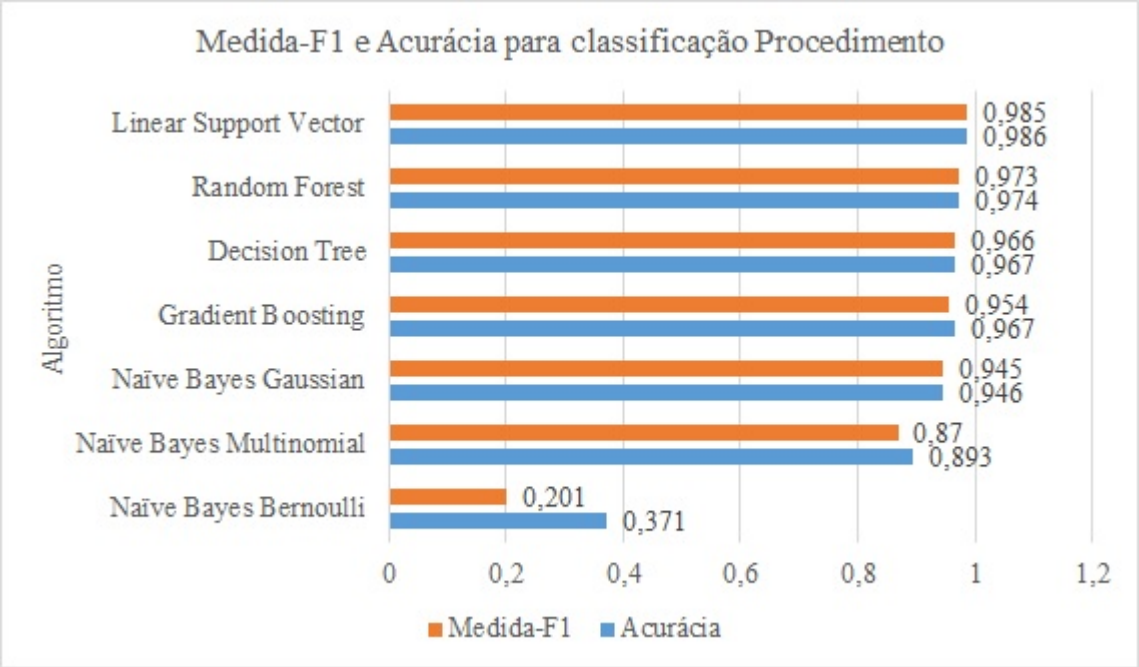
para identificar os pares significativamente diferentes: *Naïve Bayes Bernoulli - Decision Tree; Naïve Bayes Bernoulli - Naïve Bayes Gaussian; Naïve Bayes Bernoulli - Gradient Boosting; Naïve Bayes Bernoulli - Naïve Bayes Multinomial; Naïve Bayes Bernoulli - Random Forest; Naïve Bayes Bernoulli - Linear Support Vector; Decision Tree - Naïve Bayes Gaussian; Decision Tree - Naïve Bayes Multinomial; Decision Tree - Linear Support Vector; Naïve Bayes Gaussian - Gradient Boosting; Naïve Bayes Gaussian - Naïve Bayes Multinomial; Naïve Bayes Gaussian - Random Forest; Naïve Bayes Gaussian - Linear Support Vector; Gradient Boosting - Naïve Bayes*

Algoritmos	**ACU**	**SEN**	**PRE**	**F1**	**TMT**	**TMC**
Linear Support Vector	**0,986**	**0,986**	**0,985**	**0,985**	27,521 s	0,885 s
Random Forest	0,974	0,974	0,975	0,973	257,543 s	1,436 s
Decision Tree	0,967	0,967	0,968	0,966	240,776 s	**0,342 s**
Gradient Boosting	0,967	0,967	0,945	0,954	410971,485 s	1,975 s
Naïve Bayes Gaussian	0,946	0,946	0,949	0,945	14,539 s	277,073 s
Naïve Bayes Multinomial	0,893	0,893	0,860	0,870	**9,588 s**	0,732 s
Naïve Bayes Bernoulli	0,371	0,371	0,138	0,201	13,807 s	2,541 s

Tabela 13 – Comparativo das médias das métricas para classificação Procedimento.

Fonte: Autores.

Figura 13 – Métricas Medida-F1 e Acurácia para classificação Procedimento.



Fonte: Autores.

Multinomial; Gradient Boosting - Linear Support Vector; Naïve Bayes Multinomial - Random Forest; Naïve Bayes Multinomial - Linear Support Vector. A Tabela 15 apresenta os resultados do teste.

Com o classificador, também é possível gerar relatórios inteligentes que auxiliam os auditores do AudSUS em tomadas de decisão. A Figura 14 mostra um dashboard com o volume de compras de OPME realizadas pelo estado do Rio Grande do Norte (RN), o qual mostra o volume de compras por Tipo, Classe e Procedimento. Os dados do dashboard representam dados fictícios e foram construídos com a aplicação das três classificações em cada nota fiscal

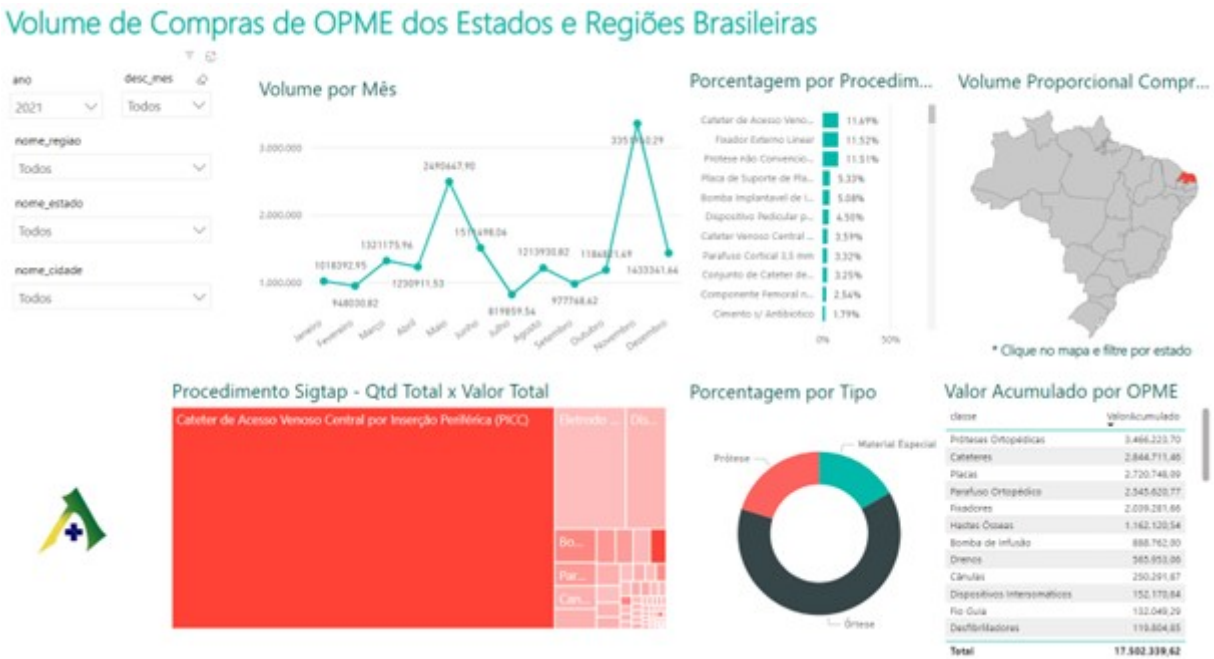
Algoritmos	**ACU**	**F1**
Decision Tree	0,811	0,331
Naïve Bayes Gaussian	0,671	0,781
Naïve Bayes Multinomial	0,464	0,384
Random Forest	0,184	0,406
Naïve Bayes Bernoulli	0,141	0,145
Linear Support Vector	0,153	0,323
Gradient Boosting	0,120	0,120

Tabela 14 – Resultado do Teste de *Shapiro-Wilk*, para análise da normalidade dos dados de Procedimento.

Fonte: Autores.

disponibilizada.

Figura 14 – Volume de Compras de OPME.



Fonte: Autores.

4.6.1 Ameaças à validade

Para validação de um experimento, é preciso considerar questões que influenciam no resultado final. Nesta seção, são apresentadas as ameaças encontradas durante a experimentação.

Comparação Dois a Dois		
Algoritmos	**ACU**	**F1**
Naïve Bayes Bernoulli - Decision Tree	0,000	0,000
Naïve Bayes Bernoulli - Naïve Bayes Gaussian	0,000	0,000
Naïve Bayes Bernoulli - Gradient Boosting	0,000	0,000
Naïve Bayes Bernoulli - Naïve Bayes Multinomial	0,000	0,000
Naïve Bayes Bernoulli - Random Forest	0,000	0,000
Naïve Bayes Bernoulli - Linear Support Vector	0,000	0,000
Decision Tree - Naïve Bayes Gaussian	0,001	0,004
Decision Tree - Gradient Boosting	1	0,334
Decision Tree - Naïve Bayes Multinomial	0,000	0,000
Decision Tree - Random Forest	0,683	0,783
Decision Tree - Linear Support Vector	0,002	0,006
Naïve Bayes Gaussian - Gradient Boosting	0,001	0,570
Naïve Bayes Gaussian - Naïve Bayes Multinomial	0,000	0,000
Naïve Bayes Gaussian - Random Forest	0,000	0,000
Naïve Bayes Gaussian - Linear Support Vector	0,000	0,000
Gradient Boosting - Naïve Bayes Multinomial	0,000	0,000
Gradient Boosting - Random Forest	0,656	0,011
Gradient Boosting - Linear Support Vector	0,002	0,000
Naïve Bayes Multinomial - Random Forest	0,000	0,000
Naïve Bayes Multinomial - Linear Support Vector	0,000	0,000
Random Forest - Linear Support Vector	0,190	0,248

Tabela 15 – Teste de *Tukey HSD*, dois a dois.

Fonte: Autores.

Validade de construção e interna: É preciso que haja uma relação de causa e efeito entre o tratamento e resultado. Os dados de notas fiscais obtidos do governo do RN foram analisados e tratados pelos autores, e assim, é preciso considerar essa ameaça. Para minimizar esse tipo de ameaça, os artefatos construídos para tratamento dos dados foram homologados e revisados por mais de um pesquisador e testes foram feitos tanto na fase de construção (validade de construção) dos artefatos, quanto na fase de execução (validade interna).

Validade externa: Apesar dos dados serem disponibilizados pelo governo do RN, não se

pode garantir que não exista algum tipo de subnotificação ou informação incorreta nos arquivos, o que pode influenciar diretamente no resultado da pesquisa.

5
Discussão

Diversas abordagens e técnicas inteligentes foram utilizadas para mitigação do problema de descrições textuais incongruentes em notas fiscais. O maior número de artigos publicados sobre o tema ocorreu em 2021, o que indica que a área de pesquisa ainda é jovem. Os resultados também mostraram que as publicações relacionadas ao tema estão presentes em vários países, o que denota que a busca por uma solução em descrições incongruentes de notas fiscais apresenta-se como uma preocupação global. A maioria dos trabalhos abordou o problema como sendo de classificação, com a ideia de agrupar os produtos, ou classificar os produtos em seu código correto. É importante ressaltar que as ferramentas, código-fonte e os detalhes de implementação não foram disponibilizados pelos autores, o que impossibilitou a verificação da eficácia das soluções propostas.

Um desafio premente consiste na escassez de estudos que busquem replicar e consolidar as pesquisas já divulgadas, além da carência de experimentos conduzidos com protocolos mais rigorosos, possibilitando tais replicabilidades. Em contrapartida, os resultados do mapeamento sistemático evidenciaram que o problema das descrições textuais incongruentes em notas fiscais é uma questão enfrentada em diversos países, permitindo uma exploração do problema com maior profundidade, inclusive considerando adaptações à realidade brasileira.

Além disso, o estudo trouxe contribuições relevantes à academia, fornecendo uma fonte de consulta com as principais propostas de solução para descrições textuais incongruentes em notas fiscais, as quais podem auxiliar as autoridades em processos de auditoria, de combate à corrupção e para o aumento da arrecadação, incrementado também os investimentos e benefícios para a população.

Após a execução do mapeamento sistemático, a criação de um modelo de conhecimento mostrou-se um processo contínuo e desafiador. Inicialmente, no experimento controlado apresentado na seção 4, o desafio estava na coleta de dados e no pré-processamento das informações necessárias para construir os modelos. Um dicionário de dados (*corpus*) foi construído com

informações de diversas fontes. Para coletar os dados do grupo 65 - 'EQUIPAMENTOS E ARTIGOS PARA USO MÉDICO, DENTÁRIO E VETERINÁRIO' do CATMAT, por exemplo, desenvolvemos um script em linguagem *Python*. No entanto, enfrentamos desafios devido à instabilidade do site do Governo Federal, o que resultou em um tempo considerável para reunir todos os dados disponíveis. Nosso script registrava o momento em que o site parava de responder e tentava periodicamente baixar os dados.

Uma peculiaridade do estudo realizado é a predominância de notas fiscais com produtos que não são OPMEs. Em certos algoritmos de aprendizado de máquina, como o *Naïve Bayes*, a decisão de estimar ou não as probabilidades *a priori* das classes com base nos dados de treinamento é crucial. Quando essas probabilidades não são calculadas, o modelo as considera uniformes, impactando a forma como o algoritmo realiza suas previsões, especialmente em conjuntos de dados com distribuições desiguais de classes, como este em questão. Por esse motivo, após testes realizados com e sem o cálculo das probabilidades *a priori* nos algoritmos, o *Naïve Bayes Multinomial* apresentou uma acurácia maior quando essas probabilidades não foram calculadas, diferindo dos demais modelos testados.

Em se tratando de desempenho, a ideia principal por trás do algoritmo *Gradient Boosting* é criar um modelo forte a partir de modelos individuais simples. Em cada iteração, o algoritmo constrói uma nova árvore de decisão que visa corrigir os erros cometidos pelas árvores anteriores. Isso é feito atribuindo pesos maiores para os exemplos que foram classificados incorretamente no modelo anterior, permitindo que o novo modelo se concentre mais nesses casos. Apesar da boa acurácia apresentada pelo algoritmo, vale destacar a dificuldade de treinamento do modelo, chegando a durar mais de quatro dias para treinar os dados para classificação Procedimento.

Com relação aos vencedores, os algoritmos *Linear Support Vector* e *Random Forest* diferem em suas abordagens. O *Linear Support Vector* procura encontrar o hiperplano que melhor separa os pontos de diferentes classes no espaço, maximizando a margem entre as classes. Já o *Random Forest* consiste em múltiplas árvores de decisão treinadas em diferentes subconjuntos do conjunto de dados, e as previsões são feitas com base na média ou voto das previsões individuais das árvores. Porém, em alguns contextos ou conjunto de dados específicos, os desempenhos dos dois algoritmos podem ser similares, como foi o caso do estudo aqui apresentado. Os algoritmos foram os que mais se destacaram em termos de acurácia para os três tipos de classificação, e, baseando-se nos testes estatísticos, apesar do *Linear Support Vector* ter uma acurácia maior, os algoritmos são semelhantes estatisticamente.

6
Conclusão

Dos 225 trabalhos recuperados das bases científicas, 15 foram aceitos pelos critérios de inclusão e exclusão, dos quais 47% foram publicados apenas em 2021. Este fato revela a tendência da área, com uma atenção recente de mais pesquisadores para o problema. Entre os principais meios de publicação, as conferências se destacaram, com 13 (87%) trabalhos, enquanto os periódicos foram representados por apenas 2 artigos (13%). Esses resultados, bem como o fato de que nenhum artigo seguiu um processo de experimentação controlada, seguindo diretrizes que permitam a realização de replicações para consolidar e validar as ferramentas, indicaram que a área ainda precisa amadurecer. As redes neurais convolucionais (CNN) e recorrentes (RNN) foram as técnicas mais utilizadas pelos autores, e entre os países, a China liderou o ranking de publicações sobre o tema.

Em se tratando do experimento controlado, este verificou as métricas de acurácia, sensibilidade, precisão e medida F-1. O destaque foi para o algoritmo *Linear Support Vector*, que obteve acima de 99% de acurácia para as classificações de Tipo e Classe, e acima de 98% de acurácia para a classificação de Procedimento. Com relação a eficiência dos algoritmos, o *Naïve Bayes Multinomial* se destacou com o tempo médio de treinamento mais rápido para os três tipos de classificação, e o *Decision Tree* foi o mais rápido em relação ao tempo médio de classificação. Essas classificações otimizarão a realização de auditorias pelo AudSUS, as quais poderão encontrar indícios tais como preços anomalamente altos, quantidades de OPMEs compradas por habitante, etc.

6.1 Contribuições

A principal contribuição deste estudo consiste na implementação e avaliação do classificador de OPMEs em notas fiscais, **OPM**in**E**r. Além disso, destacam-se as seguintes contribuições:

- Mapeamento Sistemático que teve como objetivo identificar e caracterizar as abordagens,

técnicas e algoritmos utilizados para classificação automática de descrições textuais incongruentes. Ressalta-se que os resultados foram publicados no Simpósio Brasileiro de Sistemas de Informação (SBSI 2022) (GOMES; COLAÇO JÚNIOR, 2022);

- O artigo "Prova de Conceito de um Classificador de OPMEs em Notas Fiscais", que teve por objetivo avaliar preliminarmente uma ferramenta, baseada em inteligência artificial, **OPM**in**E**r, apresentando uma prova de conceito (PoC). O artigo foi publicado no SBCAS 2023: XXIII Simpósio Brasileiro de Computação Aplicada à Saúde (GOMES et al., 2023).
- O artigo "Um Classificador Inteligente de OPMEs em Notas Fiscais", que teve por objetivo detalhar e avaliar uma ferramenta, baseada em inteligência artificial, **OPM**in**E**r, apresentando um exeperimento controlado. O artigo foi submetido à Revista Ciência & Saúde Coletiva.

6.2 Trabalhos Futuros

Além da pesquisa desenvolvida neste trabalho, outros possíveis desdobramentos são:

- Ampliar o dicionário de dados com notas fiscais de outros estados;
- Realizar experimentos com algoritmos de *deep learning*, como, por exemplo, as redes neurais convolucionais (CNN) e recorrentes (RNN), elencados no mapeamento sistemático (vide seção 2);
- No momento da publicação desta dissertação, está sendo efetuada a implementação e avaliação do algoritmo BERT (*Bidirectional Encoder Representations from Transformers*), que é um modelo de linguagem capaz de atribuir uma pontuação de probabilidade a cada palavra em uma sentença.

Referências

ABIIS. *Comércio Internacional de Produtos do Setor*. 2022. Disponível em: <https://abiis.org.br/dados-economicos-dados-economicos/>. Citado na página 15.

ALMEIDA, G. et al. Improvement of transparency through mining techniques for reclassification of texts: The case of brazilian transparency portal. In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. New York, NY, USA: Association for Computing Machinery, 2018. (dg.o '18). ISBN 9781450365260. Disponível em: <https://doi.org/10.1145/3209281.3209332>. Citado na página 31.

ALTAHERI, F.; SHAALAN, K. Exploring machine learning models to predict harmonized system code. In: THEMISTOCLEOUS, M.; PAPADAKI, M. (Ed.). *Information Systems*. Cham: Springer International Publishing, 2020. p. 291–303. ISBN 978-3-030-44322-1. Citado na página 31.

BATISTA, R.; BAGATINI, D.; FROZZA, R. Classificação automática de códigos ncm utilizando o algoritmo naïve bayes. *iSys - Brazilian Journal of Information Systems*, v. 13, p. 4–29, 06 2018. Citado 2 vezes nas páginas 16 e 17.

BERGIN, S.; WRAIGHT, P. Silver based wound dressings and topical agents for treating diabetic foot ulcers (review). *Cochrane database of systematic reviews (Online)*, v. 2006, p. CD005082, 01 2006. Citado na página 21.

BRASIL. *Constituição da República Federativa do Brasil*. 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Citado na página 15.

BRASIL. *Emenda Constitucional nº 42*. 2003. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/emendas/emc/emc42.htm>. Citado na página 37.

CAPES. *Portal de periódicos CAPES/MEC [Journal Portal CAPES/MEC]*. 2021. Disponível em: <https://www.periodicos.capes.gov.br>. Acesso em: 01 dezembro 2021. Citado na página 22.

CARVALHO, R. et al. Using clustering and text mining to create a reference price database. *Learning and Nonlinear Models*, v. 12, p. 38–52, 01 2014. Citado na página 38.

CGU, C.-G. d. U. *Portal da Transparência*. 2020. Disponível em: <http://www.portaltransparencia.gov.br/orcamento?ano=2020>. Citado na página 15.

CHAPETTA, W. A. *Uma Infra-estrutura para Planejamento, Execução e Empacotamento de Estudos Experimentais em Engenharia de Software*. Tese (Doutorado) — Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ, Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil, 2006. Citado na página 32.

CHEN, H. et al. The use of machine learning to identify the correctness of hs code for the customs import declarations. In: *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. [S.l.: s.n.], 2021. p. 1–8. Citado 2 vezes nas páginas 17 e 31.

CHEN, X.; BROMURI, S.; EEKELEN, M. van. Neural machine translation for harmonized system codes prediction. In: *2021 6th International Conference on Machine Learning*

Technologies. New York, NY, USA: Association for Computing Machinery, 2021. (ICMLT 2021), p. 158–163. ISBN 9781450389402. Disponível em: <https://doi.org/10.1145/3468891.3468915>. Citado na página 31.

COLAÇO JÚNIOR, M. et al. Evaluation of a process for the experimental development of data mining, ai and data science applications aligned with the strategic planning. *JISTEM - Journal of Information Systems and Technology Management*, TECSI Laboratório de Tecnologia e Sistemas de Informação - FEA/USP, v. 19, p. e202219018, 2022. ISSN 1807-1775. Disponível em: <https://doi.org/10.4301/S1807-1775202219018>. Citado na página 40.

CORREA, M. A. O. S.; LEAL, A. G. Identification of overpricing in the purchase of medication by the federal government of brazil, using text mining and clustering based on ontology. In: *Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing*. New York, NY, USA: Association for Computing Machinery, 2018. (ICCBDC'18), p. 66–70. ISBN 9781450364744. Disponível em: <https://doi.org/10.1145/3264560.3264569>. Citado 2 vezes nas páginas 16 e 31.

CRUZ, R.; COLAÇO JÚNIOR, M.; GOIS, V. Quão experimentais e estratégicas são as aplicações de business intelligence (bi) e data mining? ; how experimental and strategic are business intelligence (bi) and data mining applications? *Revista Ibero-Americana de Estratégia*, v. 21, p. e17689, 05 2022. Citado na página 33.

CRUZ, R. Carmo de S. et al. Análise do impacto do banco de preços em saúde (bps) para redução das assimetrias de informação dos preços de compras de Órteses, prótese e materiais especiais (opme). *JMPHC | Journal of Management amp; Primary Health Care | ISSN 2179-6750*, v. 14, n. spec, p. e006, ago. 2022. Disponível em: <https://www.jmphc.com.br/jmphc/article/view/1200>. Citado na página 16.

DING, L.; FAN, Z.; CHEN, D. Auto-categorization of hs code using background net approach. *Procedia Computer Science*, v. 60, p. 1462–1471, 12 2015. Citado na página 16.

DJANGO. *Django 4.1.7 documentation*. 2023. Disponível em: <https://docs.djangoproject.com/en/4.1>. Citado na página 45.

DU, S. et al. Hscodenet: Combining hierarchical sequential and global spatial information of text for commodity hs code classification. In: KARLAPALEM, K. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing, 2021. p. 676–689. ISBN 978-3-030-75765-6. Citado 2 vezes nas páginas 28 e 30.

DUNN, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, v. 56, p. 52–64, 1961. Disponível em: <https://api.semanticscholar.org/CorpusID:122009246>. Citado na página 46.

FEDERAL, R. *NCM*. 2023. Disponível em: <https://www.gov.br/receitafederal/pt-br/assuntos>. Citado na página 38.

FIELD, A. *Descobrindo a estatistica usando o SPSS - 2.ed.* Grupo A - Bookman, 2009. ISBN 9788536320182. Disponível em: <https://books.google.com.br/books?id=Zq059wGcnvwC>. Citado na página 46.

FIESP, F. d. I. d. E. d. S. *Corrupção: custos econômicos e propostas de combate*. 2010. Disponível em: <https://www.fiesp.com.br/arquivo-download/?id=2021>. Citado na página 15.

GOMES, W. et al. Prova de conceito de um classificador de opmes em notas fiscais. In: *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2023. p. 198–204. ISSN 2763-8987. Disponível em: <https://sol.sbc.org.br/index.php/sbcas_estendido/article/view/25355>. Citado 2 vezes nas páginas 19 e 59.

GOMES, W. F.; COLAÇO JÚNIOR, M. Applications of artificial intelligence for auditing and classification of incongruent descriptions in public procurement. In: *XVIII Brazilian Symposium on Information Systems*. New York, NY, USA: Association for Computing Machinery, 2022. (SBSI). ISBN 9781450396981. Disponível em: <https://doi.org/10.1145/3535511.3535551>. Citado 4 vezes nas páginas 19, 33, 45 e 59.

HE, M. et al. A commodity classification framework based on machine learning for analysis of trade declaration. *Symmetry*, v. 13, p. 964, 05 2021. Citado na página 30.

IBGE, I. B. d. G. e. E. *Produto Interno Bruto – PIB. 2020*. 2020. Disponível em: <https://www.ibge.gov.br/explica/pib.php>. Citado na página 15.

IBPT. *Boletim do Impostômetro mostra que ICMS tem a maior fatia de impostos recolhidos no país*. 2019. Disponível em: <https://matogrossoeconomico.com.br/economia/boletim-do-impostometro-mostra-que-icms-tem-a-maior-fatia-de-impostos-recolhidos-no-pais>. Citado na página 37.

INTERNATIONAL, T. *Corruption Perceptions Index*. 2022. Disponível em: <https://www.transparency.org/en/cpi/2022>. Citado na página 15.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele Univ.*, v. 33, 08 2004. Citado 3 vezes nas páginas 19, 21 e 24.

KRATCOSKI, P. C. Introduction: Overview of major types of fraud and corruption. In: ____. *Fraud and Corruption: Major Types, Prevention, and Control*. Cham: Springer International Publishing, 2018. p. 3–19. ISBN 978-3-319-92333-8. Disponível em: <https://doi.org/10.1007/978-3-319-92333-8_1>. Citado na página 14.

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, v. 47, p. 583–621, 1952. Disponível em: <https://api.semanticscholar.org/CorpusID:51902974>. Citado na página 46.

MANKIW, N. G.; ROMER, D.; WEIL, D. N. A Contribution to the Empirics of Economic Growth*. *The Quarterly Journal of Economics*, v. 107, n. 2, p. 407–437, 05 1992. ISSN 0033-5533. Disponível em: <https://doi.org/10.2307/2118477>. Citado na página 15.

MLFLOW. *MLflow 2.3 documentation*. 2023. Disponível em: <https://www.mlflow.org/docs/2.3>. Citado na página 45.

MS, M. d. S. *Ministério da Saúde*. 2017. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2017/prt3992_28_12_2017.html>. Citado na página 15.

MUIR, D. R. W. A. Using machine learning to improve public reporting on u.s. government contracts. *INFORMS Journal on Applied Analytics*, v. 51, n. 6, p. 463–479, 2021. Citado na página 31.

NATIONS, U. *United Nations Convention Against Corruption.* 2004. Disponível em: <https://www.unodc.org/documents/brussels/UN_Convention_Against_Corruption.pdf>. Citado na página 14.

PAIVA, E. de; REVOREDO, K. Big data e transparência: Utilizando funções de mapreduce para incrementar a transparência dos gastos públicos. In: *Anais do XII Simpósio Brasileiro de Sistemas de Informação*. Porto Alegre, RS, Brasil: SBC, 2016. p. 025–032. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/5942>. Citado na página 16.

PARAMARTHA, I. G. Y.; ARDIYANTO, I.; HIDAYAT, R. Developing machine learning framework to classify harmonized system code. case study: Indonesian customs. In: *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*. [S.l.: s.n.], 2021. p. 254–259. Citado na página 31.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, 01 2012. Citado na página 45.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, v. 64, 08 2015. Citado na página 21.

POSTGRESQL. *PostgreSQL 12.12 documentation*. 2021. Disponível em: <https://www.postgresql.org/docs/12>. Citado na página 45.

POWERBI. *Power BI Desktop (Version 2.117.984.0)*. 2023. Disponível em: <https://powerbi.microsoft.com>. Citado na página 45.

PYTHON. *Python 3.10*. 2021. Disponível em: <https://www.python.org>. Citado na página 45.

RIBEIRO, L. et al. Reconhecimento de entidades nomeadas em itens de produto da nota fiscal eletrônica. v. 36, p. 116–126, 01 2018. Citado na página 16.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, v. 24, p. 513–523, 1988. Disponível em: <https://api.semanticscholar.org/CorpusID:7725217>. Citado na página 38.

SALTON, G.; FOX, E. A.; WU, H. Extended boolean information retrieval. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 26, n. 11, p. 1022–1036, nov 1983. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/182.358466>. Citado na página 38.

SANTOS, B. et al. Análise comparativa de algoritmos de mineração de texto aplicados a históricos de contas públicas. In: *Anais do XI Simpósio Brasileiro de Sistemas de Informação*. Porto Alegre, RS, Brasil: SBC, 2015. p. 667–674. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/5874>. Citado na página 33.

SANTOS, C. M. d. C.; PIMENTA, C. A. d. M.; NOBRE, M. R. C. A estratégia pico para a construção da pergunta de pesquisa e busca de evidências. *Revista Latino-Americana de Enfermagem*, v. 15, n. 3, p. 508–511, jun. 2007. Disponível em: <https://www.revistas.usp.br/rlae/article/view/2463>. Citado na página 21.

SCOPUS. *Scopus - Elsevier Database*. 2021. Disponível em: <https://www.scopus.com>. Acesso em: 01 dezembro 2021. Citado na página 24.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, v. 52, p. 591–611, 1965. Disponível em: <https://api.semanticscholar.org/CorpusID:124868013>. Citado na página 46.

SPICHAKOVA, M.; HAAV, H.-M. Application of machine learning for assessment of hs code correctness. *Balt. J. Mod. Comput.*, v. 8, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:231876846>. Citado 2 vezes nas páginas 17 e 31.

TCU. *Auditoria em òrtese, Prótese e Materiais Especiais (OPME)*. 2016. Disponível em: <https://portal.tcu.gov.br/biblioteca-digital/auditoria-em-ortese-protese-e-materiais-especiais-opme.htm>. Citado na página 16.

TRAVASSOS, G. *Introdução à engenharia de software experimental*. UFRJ, 2002. (RT-ES-590/02). Disponível em: <https://books.google.com.br/books?id=4SnKZwEACAAJ>. Citado na página 41.

VIEIRA F. S.; PIOLA, S. F. B. R. P. d. S. e. *Vinculação orçamentária do gasto em saúde no Brasil: resultados e argumentos a seu favor*. 2019. Disponível em: <https://saudeamanha.fiocruz.br/wp-content/uploads/2019/10/td_2516.pdf>. Citado na página 15.

VIJAYARANI, S.; JANANI, M. Text mining: open source tokenization tools – an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, v. 3, 01 2016. Citado na página 38.

WOHLIN, C. et al. Experimentation in software engineering. In: *The Kluwer International Series in Software Engineering*. [s.n.], 2000. Disponível em: <https://api.semanticscholar.org/CorpusID:9558494>. Citado na página 19.

YU, J. et al. Neural network based transaction classification system for chinese transaction behavior analysis. In: *2019 IEEE International Congress on Big Data (BigDataCongress)*. [S.l.: s.n.], 2019. p. 64–71. Citado na página 30.

YUE, Y. et al. Extremely short chinese text classification method based on bidirectional semantic extension. *Journal of Physics: Conference Series*, v. 1437, p. 012026, 01 2020. Citado 2 vezes nas páginas 28 e 31.

ZHU, W.; ZENG, N.; WANG, N. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas ® implementations. *NorthEast SAS users group, health care and life sciences*, 01 2010. Citado na página 39.