



UNIVERSIDADE FEDERAL DO MARANHÃO
Programa de Pós-Graduação em Ciência da Computação

Igor Bezerra Reis

***Um Método Baseado em Grandes Modelos de
Linguagem Generativos para a Análise de
Similaridades entre Marcas no Contexto
Brasileiro***

São Luís

2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO MARANHÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**Um Método Baseado em Grandes Modelos de
Linguagem Generativos para a Análise de
Similaridades entre Marcas no Contexto
Brasileiro**

Igor Bezerra Reis

São Luís, 2025

Igor Bezerra Reis

**Um Método Baseado em Grandes Modelos de Linguagem
Generativos para a Análise de Similaridades entre Marcas
no Contexto Brasileiro**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Maranhão, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ariel Soares Teles
UFMA/IFMA

São Luís

2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Bezerra Reis, Igor.

Um Método Baseado em Grandes Modelos de Linguagem
Generativos para a Análise de Similaridades entre Marcas
no Contexto Brasileiro / Igor Bezerra Reis. - 2025.
99 f.

Orientador(a): Ariel Soares Teles.

Dissertação (Mestrado) - Programa de Pós-graduação em
Ciência da Computação/ccet, Universidade Federal do
Maranhão, São Luís, 2025.

1. Marcas. 2. Propriedade Intelectual. 3.
Similaridade de Marcas. 4. Modelos de Linguagem de Grande
Escala. I. Soares Teles, Ariel. II. Título.

IGOR BEZERRA REIS

**UM MÉTODO BASEADO EM GRANDES MODELOS DE LINGUAGEM
GENERATIVOS PARA A ANÁLISE DE SIMILARIDADES ENTRE
MARCAS NO CONTEXTO BRASILEIRO**

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

Área de Concentração: Ciência da Computação, Propriedade Intelectual.

Linha de Pesquisa: Inteligência Artificial aplicada à Propriedade Intelectual.

Aprovada em 22/04/2025.

COMISSÃO EXAMINADORA

Prof. Dr. Ariel Soares Teles

Orientador

Universidade Federal do Maranhão / Instituto Federal do Maranhão

Prof. Dr. Luciano Reis Coutinho - 1º Membro

Examinador Interno

Universidade Federal do Maranhão

Prof. Dr. Cícero Eduardo de Sousa Walter – 2º Membro

Examinador Externo

Instituto Federal do Piauí

Dedicatória: Ao meu pai, minha mãe e meu irmão, que me deram toda a base necessária para que eu pudesse estar aqui. Ao Lupi e Beethoven. Nada disso faria sentido se não fosse pelo desejo profundo de deixá-los orgulhosos. Esta conquista é, acima de tudo, de vocês.

Agradecimentos

Em primeiro lugar, gostaria de agradecer à minha família: minha mãe, Maria Célia Alves Bezerra Reis; meu pai, Antão Almeida de Jesus Reis; e meu irmão, Hyago Bezerra Reis, que sempre foram meu apoio em todos os momentos da vida. Saibam que, se cheguei até aqui — e até onde ainda irei chegar, pela graça de Deus —, vocês serão os grandes responsáveis. A vocês dedico todo o meu amor, suor, dedicação e carinho. Obrigado por sempre me incentivarem a estudar e a seguir o bom caminho, por terem enfrentado as maiores dificuldades da vida para que eu pudesse realizar meus sonhos. Se alcancei a sombra, foi porque vocês caminharam sob o sol.

Ao meu cachorro Lupi, por tantas noites em que me ouviu falar sobre meus sonhos e medos. O carregarei comigo por toda a vida.

Ao meu Deus e Senhor Jesus Cristo, que jamais me abandonou e que, nas noites de dúvida, tristeza ou alegria, sempre esteve ao meu lado.

Ao Prof. Dr. Ariel Soares Teles, que me orientou com dedicação ao longo de toda esta pesquisa. Agradeço profundamente pela confiança depositada em meu trabalho desde nossa primeira reunião, quando decidiu abraçar este tema de pesquisa, até a conclusão desta etapa tão significativa da minha vida. Sua paciência constante, generosidade e ensinamentos valiosos foram essenciais para o meu crescimento acadêmico e pessoal. Serei eternamente grato por sua orientação atenta e por me ensinar não apenas a conduzir este trabalho, mas também a trilhar o caminho da pesquisa com paixão e comprometimento. Tenho por Vossa Senhoria profunda admiração e respeito.

Ao meu grande amigo, que também se tornou um verdadeiro irmão, Bruno Roberto Silva de Moraes, por me acolher com tanto carinho em São Luís. Saiba que, sem a sua ajuda, eu não teria conseguido trilhar este caminho. Muito obrigado, meu irmão, por toda a ajuda, apoio, conversas e conselhos que sempre levarei comigo com gratidão.

Ao meu grande amigo da vida, Rafael Ângelo Santos Leite, que começou como meu orientador na graduação e logo se tornou um amigo para além da academia. Agradeço por sempre acreditar em mim e por me incentivar a seguir em frente, mesmo nos momentos mais difíceis. Obrigado por cada orientação, conversa e conselho ao longo das etapas da minha vida. Grande parte desta conquista também é sua. Sou imensamente grato a Deus por sua amizade.

A todos os meus amigos que, de forma direta ou indireta, caminharam comigo ao longo desta jornada, em especial Letícia Alves Lemos e Fátima Milena de Sousa Alves, por sempre me apoiarem — desde as vezes em que me acompanharam à rodoviária até

as conversas mais profundas, seja à distância ou presencialmente. À minha amiga Lara Estéfany Reis de Sousa, por todas as conversas, conselhos, incentivos e preocupações constantes. Aos meus grandes amigos Erick Pereira Silva, Marcos Sérgio da Luz Santos e Ivisson da Silva Vieira, pelas conversas, diálogos e reuniões de todo final de semana. Sou imensamente grato por compreenderem minhas dores, angústias e anseios, e por estarem sempre presentes nos momentos em que mais precisei. Vocês tornaram os dias mais leves e os momentos difíceis mais suportáveis. Sem dúvida, me ajudaram a enxergar o sentido das coisas e a continuar com coragem nesta caminhada. Obrigado por compreenderem minha ausência e por darem significado a ela.

Aos professores membros da banca examinadora — Prof. Dr. Ariel Soares Teles, Prof. Dr. Luciano Reis Coutinho e Prof. Dr. Cícero Eduardo de Sousa Walter — agradeço pela generosidade em dedicar seu tempo à leitura cuidadosa deste trabalho e pelas valiosas sugestões e contribuições, que enriqueceram significativamente esta pesquisa.

Agradeço ao Laboratório de Sistemas Distribuídos Inteligentes (LSDi) da UFMA pela disponibilização do equipamento computacional utilizado na realização de todos os experimentos, viabilizado por meio do EDITAL FAPEMA/EMAP nº 12/2022 — Apoio a Pesquisas no Porto do Itaqui, projeto APP-09405/22.

Expresso minha gratidão à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [Código de Financiamento 001] e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [bolsa 308059/2022-0] pelo apoio financeiro essencial à realização desta pesquisa.

Agradeço ainda a todos os docentes que contribuíram para a minha formação acadêmica, especialmente aos professores do Programa de Pós-Graduação em Ciência da Computação (PPGCC/UFMA), pelos ensinamentos e pela excelência na condução das disciplinas do curso.

You're on Your Own, Kid. You always have been. Long story short, I survived.

“E ainda que tivesse o dom de profecia, e conhecesse todos os mistérios e todo o conhecimento, e ainda que tivesse toda a fé, de maneira tal que transportasse os montes, e não tivesse amor, nada seria.” 1 Coríntios 13.

Resumo

Uma marca tem o objetivo de identificar, de forma única e distintiva, os produtos e serviços oferecidos por uma empresa. Ela constitui um importante ativo intangível, atuando como instrumento fundamental para impedir práticas de concorrência desleal e fortalecer o posicionamento competitivo das empresas no mercado. Contudo, o crescente número de pedidos de registro de marcas submetidos ao Instituto Nacional de Propriedade Intelectual (INPI) tem gerado desafios significativos, incluindo aumento no tempo de análise, inconsistências nas decisões e maior complexidade na identificação de conflitos entre marcas. Diante desse cenário, métodos automatizados para a análise de similaridade entre marcas tornam-se essenciais para aprimorar a eficiência, confiabilidade e agilidade dos processos realizados pelo instituto. Neste contexto, esta dissertação propõe um método baseado em Modelos de Linguagem de Grande Escala (LLMs) para classificar e explicar a similaridade entre marcas nominativas, considerando os critérios fonético, ideológico, distintivo e mercadológico adotados pelo INPI. Para alcançar esse objetivo, o método foi estruturado em duas etapas principais: (1) desenvolvimento de um modelo classificador, capaz de identificar rapidamente e com precisão conflitos entre marcas; e (2) desenvolvimento de um modelo explicador, responsável por fornecer explicações detalhadas sobre os motivos pelos quais duas marcas são consideradas similares ou não. Para desenvolver este método proposto, foi criado um Conjunto de Dados de Arquivo de Marcas Conflitantes (do inglês, *Conflicting Marks Archive Dataset - CMAD*), composto por casos reais de conflitos entre marcas extraídos das publicações oficiais do INPI. Diferentes LLMs foram avaliados quanto à precisão na classificação dos conflitos entre marcas e na geração de explicações fundamentadas sobre os casos analisados, sendo essas avaliações validadas por especialistas em Propriedade Intelectual (PI). Os resultados demonstraram que alguns dos modelos avaliados obtiveram desempenho expressivo, com acurácia próxima a 99%, F1-score superior a 98% e AUC acima de 99% na classificação dos conflitos. Além disso, as explicações fornecidas pelos modelos foram avaliadas por especialistas com média superior a 4, em uma escala de 0 a 5, demonstrando clareza e alinhamento com os critérios estabelecidos pelo INPI. Portanto, este estudo apresenta uma contribuição para a análise automatizada de similaridade entre marcas, destacando o potencial dos LLMs na otimização e aprimoramento dos processos relacionados à PI, reduzindo a subjetividade inerente à análise manual e ampliando a transparência nas decisões institucionais, possibilitando democratizar o acesso à proteção de marcas, especialmente beneficiando pequenas e médias empresas que enfrentam limitações financeiras ou técnicas, contribuindo diretamente para a redução de custos operacionais, prevenção de litígios de marcas e aceleração dos processos decisórios.

Palavras-chave: Marcas, Propriedade Intelectual, Similaridade de Marcas, Modelos de Linguagem de Grande Escala.

Abstract

A trademark aims to uniquely and distinctively identify the products and services offered by a company. It constitutes an important intangible asset, serving as a fundamental tool to prevent unfair competition practices and strengthen the company's competitive position in the market. However, the increasing number of trademark registration applications submitted to the Brazilian National Institute of Industrial Property (INPI) has posed significant challenges, including longer review times, inconsistencies in decisions, and increased complexity in identifying trademark conflicts. In this context, automated methods for trademark similarity analysis have become essential to enhance the efficiency, reliability, and speed of the institute's processes. This dissertation proposes a method based on Large Language Models (LLMs) to classify and explain the similarity between word marks, taking into account the phonetic, ideological, distinctive, and market-related criteria adopted by INPI. To achieve this goal, the method was structured into two main stages: (1) the development of a classification model capable of quickly and accurately identifying conflicts between trademarks; and (2) the development of an explanation model responsible for providing detailed justifications regarding why two trademarks are considered similar or not. To support this approach, a dataset named the Conflicting Marks Archive Dataset (CMAD) was created, comprising real trademark conflict cases extracted from INPI's official publications. Several LLMs were evaluated for their accuracy in classifying conflicts and for generating well-founded explanations of the analyzed cases, which were validated by Intellectual Property (IP) specialists. The results showed that some of the evaluated models achieved outstanding performance, with accuracy close to 99%, F1-score above 98%, and AUC above 99% in the classification task. Furthermore, the explanations provided by the models received average ratings above 4 (on a scale from 0 to 5) from experts, demonstrating clarity and alignment with the criteria established by INPI. Therefore, this study contributes to the field of automated trademark similarity analysis by highlighting the potential of LLMs to optimize and enhance processes related to IP, reducing the subjectivity inherent to manual evaluation and increasing transparency in institutional decisions. The proposed method also facilitates broader access to trademark protection, especially for small and medium-sized enterprises facing financial or technical limitations, thus helping reduce operational costs, prevent trademark disputes, and accelerate decision-making processes.

Keywords: Trademarks, Intellectual Property, Trademark Similarity, Large Language Models.

Lista de ilustrações

Figura 1 – Exemplos de Marcas Nominativas. Fonte: (INPI, 2023b)	22
Figura 2 – Exemplos de Marcas Figurativas. Fonte: (INPI, 2023b)	23
Figura 3 – Exemplo de marcas visualmente semelhantes (INPI, 2023b).	27
Figura 4 – Estrutura do mecanismo de <i>attention</i> utilizado na arquitetura <i>Transformer</i> . Fonte: Niu, Zhong e Yu (2021)	31
Figura 5 – Etapas da metodologia.	41
Figura 6 – Exemplo de indeferimento.	41
Figura 7 – Exemplo de um arquivo XML contendo um pedido de registro de marca.	42
Figura 8 – Exemplo de dados coletados de uma marca indeferida.	43
Figura 9 – Distribuição das amostras quanto à forma de apresentação: (a) pedidos de marcas indeferidas; e (b) marcas registradas.	43
Figura 10 – Três amostras de marcas colidentes no CMAD: (a) três pares de imagens de marcas; e (b) respectivas entradas no arquivo CSV.	45
Figura 11 – Visão Geral da Metodologia.	47
Figura 12 – Divisão da base de dados.	53
Figura 13 – Tela de avaliação do relatório.	55
Figura 14 – Matrizes de confusão dos modelos avaliados.	57
Figura 15 – Análise comparativa das curvas ROC entre todos os modelos (a) e os cinco com melhor desempenho (b).	59
Figura 16 – Boxplot das Notas com Média Geral por Modelo.	62
Figura 17 – Boxplot das Notas por Modelo e Tipo de Relatório.	63
Figura 18 – Comentários dos especialistas que ressaltam a importância da contextualização dos serviços associados às classes NICE para uma análise mais precisa da similaridade entre marcas.	65
Figura 19 – Comentários apontando limitações dos modelos na identificação dos elementos realmente distintivos das marcas.	66
Figura 20 – Comentários que destacam positivamente a estrutura lógica e coerência dos relatórios gerados pelos modelos.	66
Figura 21 – Comentário feito para relatório do Mistral-Nemo.	66

Lista de tabelas

Tabela 1 – Comparação entre trabalhos relacionados na avaliação de similaridade entre marcas.	36
Tabela 2 – Análise comparativa entre conjuntos de dados relacionados e o conjunto de dados desenvolvido neste trabalho.	40
Tabela 3 – Descrição dos dados do arquivo CSV do CMAD.	44
Tabela 4 – Modelos de LLMs selecionados para avaliação	51
Tabela 5 – Matriz de confusão.	53
Tabela 6 – Métricas de desempenho dos modelos avaliados, ordenadas pelo F1-Score.	58
Tabela 7 – Estatísticas das Avaliações dos Modelos.	62
Tabela 8 – Erros cometidos pelo modelo: comparação entre marcas, com rótulo e predição.	68

Lista de Siglas

ABPI Associação Brasileira da Propriedade Intelectual.

AM Aprendizado de Máquina.

AP Aprendizado Profundo.

AUC Area Under the Curve.

BERT *Bidirectional Encoder Representations from Transformers.*

CIPO Canadian Intellectual Property Office.

CMAD Conflicting Marks Archive Dataset: A Dataset of Conflicting Marks from the Brazilian Intellectual Property Office.

CNIPO China National Intellectual Property Administration.

CNNs Redes Neurais Convolucionais.

CSV Comma-Separated Values.

EPP Empresa de Pequeno Porte.

EUIPO European Union Intellectual Property Office.

FORTEC Fórum Nacional de Gestores de Inovação e Transferência de Tecnologia.

GPU Graphics Processing Unit.

IA Inteligência Artificial.

INPI Instituto Nacional da Propriedade Industrial.

IQR Intervalo Interquartil.

LLM *Large Language Models.*

LLMs Grandes Modelos de Linguagens.

LoRA Low-Rank Adaptation.

LSTM Long Short-Term Memory.

ME Microempresa.

MEI Microempreendedor Individual.

MLM Modelagem de Linguagem Mascarada.

MR Marca Registrada.

PEFT Parameter-Efficient Fine-Tuning.

PI Propriedade Intelectual.

PLN Processamento de Linguagem Natural.

PNG Portable Network Graphic.

QLoRA Efficient Finetuning of Quantized LLMs.

RLHF Reinforcement Learning from Human Feedback.

RNN Redes Neurais Recorrentes.

ROC Receiver Operating Characteristic.

RPI Revista da Propriedade Industrial.

STS Similaridade Textual Semântica.

TPU Tensor Processing Unit.

USPTO United States Patent and Trademark Office.

WIPO World Intellectual Property Organization.

XML Extensible Markup Language.

Sumário

1	INTRODUÇÃO	15
1.1	Contexto de Pesquisa	15
1.2	Caracterização do Problema	16
1.3	Relevância do Trabalho	17
1.4	Hipótese de Pesquisa	19
1.5	Objetivos	20
1.5.1	Geral	20
1.5.2	Específicos	20
1.6	Organização do Trabalho	20
2	FUNDAMENTAÇÃO TEÓRICA	22
2.1	Marcas	22
2.1.1	Aspectos Legais das Marcas no Brasil e Processo de registro	24
2.1.2	Similaridade Fonética e Textual	25
2.1.3	Similaridade Ideológica	26
2.1.4	Similaridade Visual	26
2.2	Grandes Modelos de Linguagens (LLMs)	27
2.2.1	Fine-tuning de Modelos de Linguagem de Grande Escala (LLMs)	28
2.2.2	Arquitetura dos LLMs	30
3	TRABALHOS RELACIONADOS	33
3.1	Soluções baseadas em LLMs para similaridade textual	33
3.2	Soluções para Similaridade Fonética e Textual de Marcas	34
3.3	Contribuições desta Pesquisa	35
4	CONFLICTING MARKS ARCHIVE DATASET	38
4.1	Conjunto de Dados Relacionados	38
4.2	Métodos	41
4.3	Descrição dos Dados	42
4.4	Discussão	44
4.4.1	Aplicações do Conflicting Marks Archive Dataset: A Dataset of Conflicting Marks from the Brazilian Intellectual Property Office (CMAD)	44
4.4.2	Considerações Éticas	46
4.4.3	Forças e Limitações	46
5	MATERIAIS E MÉTODOS	47

5.1	Elaboração dos prompts	48
5.2	Fine-tuning e de Avaliação dos modelos de LLMs classificadores	50
5.3	Avaliação por Especialista	54
5.4	Engenharia de Prompt com Abordagem Few-Shot: Geração e Avaliação de Relatórios por Especialistas	55
6	RESULTADOS E DISCUSSÃO	56
6.1	Modelos Classificadores	56
6.2	Modelos Explicadores	61
6.2.1	Comentários dos Especialistas	63
6.3	Discussão	64
6.3.1	Modelos Classificadores para Identificação de Similaridade	65
6.3.1.1	Análise do Desempenho dos Modelos	65
6.3.1.2	Análise de Erros	67
6.3.2	Modelos Explicadores para Geração de Relatórios	70
7	CONSIDERAÇÕES FINAIS	78
7.1	Contribuições Científicas, Tecnológicas e Sociais	80
7.2	Limitações	81
7.3	Trabalhos Futuros	82
7.4	Escrita e Publicações de Artigos Científicos	84
	REFERÊNCIAS	85
	APÊNDICES	92
	Apêndices	93
	Apêndice A Exemplo de Prompt Utilizado	93

1 Introdução

1.1 Contexto de Pesquisa

Uma Marca Registrada (MR) é um ativo intangível que concede ao seu titular o direito exclusivo de uso, impedindo que terceiros explorem indevidamente sua reputação e identidade no mercado (TRAPPEY; TRAPPEY; LIN, 2020). Além de garantir proteção legal, o registro de marca contribui para a fidelização dos clientes e o aumento das vendas, uma vez que os consumidores frequentemente associam a reputação da marca à qualidade do produto ou serviço (VESNIN; LEVSHUN; CHECHULIN, 2023). Dessa forma, a proteção das MR é essencial para evitar prejuízos financeiros decorrentes de infrações à Propriedade Intelectual (PI).

A definição de marca registrada pode variar ligeiramente entre os países, mas há um consenso global sobre sua função. Segundo o Instituto Nacional da Propriedade Industrial (INPI), no Brasil, uma marca tem como objetivos fundamentais identificar a origem e distinguir produtos ou serviços de outros semelhantes ou relacionados, mas provenientes de fontes diferentes, podendo ser classificada em quatro tipos principais: nominativa, figurativa, mista e tridimensional (INPI, 2023b). O Canadian Intellectual Property Office (CIPO) (CIPO, 1976), por sua vez, define a marca registrada como uma combinação de letras, palavras, sons ou desenhos usada para diferenciar os produtos ou serviços de uma empresa dos de outras no mercado, promovendo reconhecimento e fidelidade entre os consumidores. Já o United States Patent and Trademark Office (USPTO) (USPTO, 2024) reforça que o objetivo central das marcas registradas é permitir que os clientes identifiquem e diferenciem os produtos ou serviços de uma empresa dos de seus concorrentes. A World Intellectual Property Organization (WIPO) (OMPI,) adota uma definição abrangente, considerando a marca registrada como qualquer sinal que possa distinguir os produtos ou serviços de uma empresa. Dada a importância estratégica das MRs, o processo de avaliação da possibilidade de registro desempenha um papel essencial na prevenção de conflitos comerciais e jurídicos. No Brasil, essa responsabilidade compete ao INPI, instituição que verifica se as marcas solicitadas atendem aos critérios legais de registro, incluindo originalidade, distintividade e ausência de colidência com marcas previamente registradas (INPI, 2023b). Quando duas marcas possuem similaridades suficientes para causar confusão ou associação indevida por parte dos consumidores, ocorre o que o INPI denomina de *litígio*, oposição a um pedido de registro de marca ou um processo administrativo de nulidade.

Esse fenômeno de litígio pode acontecer devido a semelhanças em diferentes dimensões, como fonética, visual, conceitual (ideológica) ou mesmo pela proximidade mercadológica. Por exemplo, marcas que possuem nomes semelhantes ou que soam parecidas

podem induzir o consumidor a acreditar, equivocadamente, que os produtos ou serviços têm a mesma origem comercial. Essa confusão não apenas prejudica os consumidores, como também pode causar danos à reputação e perdas econômicas significativas para as empresas envolvidas (VESNIN; LEVSHUN; CHECHULIN, 2023; TRAPPEY; TRAPPEY; LIN, 2020).

A análise de similaridade entre marcas nominativas, aquelas compostas exclusivamente por palavras ou expressões sem elementos gráficos adicionais, é especialmente complexa. Isso ocorre porque, ao contrário das marcas figurativas ou mistas, que possuem elementos visuais adicionais que facilitam a diferenciação, as marcas nominativas dependem unicamente da percepção fonética e semântica do consumidor para sua distinção (INPI, 2023b). Portanto, o exame dessas marcas exige uma abordagem detalhada e criteriosa, considerando nuances linguísticas e mercadológicas para minimizar riscos de litígios.

1.2 Caracterização do Problema

Durante esse processo de análise de semelhança entre duas marcas, são considerados três principais tipos de similaridade: similaridade visual, que refere-se a análise de elementos gráficos, como logotipos, formas geométricas, cores e tipografia; similaridade fonética e textual, focada na avaliação das palavras, sequências silábicas, entonação e ritmo das frases; e similaridade ideológica, que é a análise da evocação de conceitos ou significados semelhantes, mesmo que os elementos gráficos ou fonéticos sejam distintos. A análise dessas características permite identificar potenciais conflitos e garantir que a marca possua distintividade suficiente para evitar litígios e proteger sua identidade no mercado (INPI, 2023b; INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL,).

Quando há conflitos entre marcas, essa situação frequentemente leva a litígios judiciais dispendiosos, nos quais empresas disputam o direito sobre o nome ou símbolo. O impacto financeiro dessas disputas pode ser substancial, afetando tanto pequenas empresas quanto grandes corporações (TRAPPEY; TRAPPEY; LIN, 2020). A demora no processo de análise de similaridade pode ser outro problema, o INPI recebe anualmente centenas de milhares de pedidos de registro, tornando o tempo de processamento um fator crítico para empresas que dependem da proteção de sua marca para operar no mercado, onde o processo de registro não se torna apenas complexo e oneroso, mas também se torna cada vez mais desafiador diante do aumento do volume de pedidos de registro.

Em 2018, foram realizados aproximadamente 10,9 milhões de registros de marcas em todo o mundo, representando um crescimento de 19,2% em relação a 2017 (WIPO, 2019). No Brasil, o número acumulado de pedidos de registro de marcas em 2023 aumentou 0,9% em comparação a 2022, totalizando 402.460 solicitações (INPI, 2023a). Já em 2024, o INPI registrou um total de 444.037 solicitações de marcas no Brasil, representando um

aumento de 10,3% em relação a 2023 (INDUSTRIAL, 2025). Logo, aumento contínuo do volume de registros resulta em filas extensas e prazos prolongados para a avaliação de cada pedido, percebe-se assim que a demanda por análise de colidência entre marcas cresce em ritmo acelerado, exigindo soluções mais ágeis e eficientes por parte das instituições responsáveis.

Outro aspecto problemático do atual modelo de análise de marcas é a subjetividade do processo. O exame de similaridade é conduzido por especialistas que, mesmo seguindo diretrizes estabelecidas, interpretam as características das marcas com base em sua experiência e julgamento pessoal. Essa subjetividade pode levar a decisões inconsistentes, em que marcas com características similares são aprovadas em alguns casos e rejeitadas em outros, logo, essa variabilidade, aliada ao volume crescente de registros, torna a automação do processo uma alternativa promissora (ALSHOWAISH; AL-OHALI; AL-NAFJAN, 2022), afinal, a similaridade com marcas já registradas continua sendo uma das principais razões para a rejeição de novos pedidos (SHOWKATRAMANI *et al.*, 2018a).

Assim, a caracterização do problema evidencia a necessidade urgente de aprimorar o processo de registro de marcas, reduzindo sua subjetividade e tornando a análise mais ágil e precisa. A integração de soluções baseadas em Inteligência Artificial (IA) tem o potencial de transformar esse cenário, proporcionando um sistema mais justo, transparente e alinhado às demandas do mercado moderno.

1.3 Relevância do Trabalho

Esse contexto evidencia a importância do desenvolvimento de métodos mais eficazes e precisos para apoiar o processo de avaliação realizado pelo INPI. Nesse sentido, técnicas baseadas em IA, especialmente as relacionadas a *Large Language Models* (LLM), emergem como alternativas promissoras devido à sua capacidade de processamento linguístico avançado e interpretação de contextos implícitos. Os LLM têm demonstrado grande potencial em capturar nuances sutis na linguagem, o que pode ser extremamente benéfico na avaliação de similaridade entre marcas, contribuindo para decisões mais consistentes e objetivas (BROWN *et al.*, 2020).

Modelos como GPT-4, Mistral, Gemma e LLaMA demonstraram notável versatilidade, destacando-se em tarefas como tradução automática, sumarização de texto, respostas a perguntas e até mesmo resolução de problemas complexos por meio do raciocínio em cadeia de pensamento (do inglês, *chain of thought reasoning*) (MINAEE *et al.*, 2024). Além de seu promissor desempenho na classificação de textos, esses modelos também são capazes de gerar explicações sobre suas decisões, tornando-se ferramentas valiosas para contextos onde a transparência e a interpretabilidade são fundamentais (BOGAERT *et al.*, 2023). A combinação de previsões precisas com justificativas detalhadas é essencial em

áreas onde a confiança nas decisões automatizadas é essencial (YANG *et al.*, 2024b; HU *et al.*, 2024). Isso faz dos LLMs uma solução promissora para problemas que exigem análises detalhadas e justificativas consistentes, como a verificação de similaridade entre marcas.

Em nível nacional, o Brasil registrou um aumento significativo nos pedidos de registro de marcas. Em 2023, o INPI contabilizou 402.460 solicitações, representando um crescimento de 0,9% em relação ao ano anterior. Destaca-se que 51% desses pedidos foram realizados por Microempreendedor Individual (MEI), Microempresa (ME) e Empresa de Pequeno Porte (EPP), evidenciando o reconhecimento crescente da importância da proteção de marcas (Governo do Brasil, 2023; Agência Sebrae de Notícias, 2023). Ainda em 2023, foram concedidos 210.987 registros, um aumento expressivo de 29,3% em comparação ao ano anterior (Governo do Brasil, 2023).

No cenário global, o número de pedidos de registro de marcas também tem apresentado variações expressivas. Em 2017, aproximadamente 12,39 milhões de pedidos foram registrados mundialmente, indicando um crescimento de 26,8% em relação a 2016. A China liderou esse crescimento com 5,7 milhões de pedidos, seguida por Estados Unidos, Japão, União Europeia e Índia (Organização Mundial da Propriedade Intelectual (WIPO), b). Contudo, em 2022, observou-se um declínio de 6,1% nos pedidos internacionais, totalizando cerca de 69.000 solicitações, refletindo uma desaceleração após um crescimento expressivo de 15% em 2021 (Organização Mundial da Propriedade Intelectual (WIPO), a). Essas variações reforçam a necessidade de metodologias eficazes para lidar com a crescente complexidade na gestão de registros de marcas.

Diante desse cenário, esta pesquisa investiga o uso de LLM de código aberto especializados, treinados por meio de aprendizado supervisionado, para automatizar e aprimorar a análise de similaridade entre marcas nominativas no Brasil. Esses modelos são adaptados para avaliar a similaridade conforme os critérios estabelecidos pelo INPI, incorporando aspectos fonéticos, ideológicos e distintivos de maneira estruturada. Adicionalmente, a implementação de um modelo interpretável permite fornecer justificativas detalhadas para as decisões tomadas, garantindo maior transparência e confiabilidade nas análises realizadas (MINAEE *et al.*, 2024).

A aplicação dessa abordagem pode transformar significativamente o processo de busca por anterioridade e registro de marcas, reduzindo o tempo necessário para as análises e minimizando os custos operacionais. Escritórios especializados em PI, instituições como o próprio INPI e empresas de diversos portes poderão realizar buscas mais rápidas e precisas, facilitando a triagem antecipada e identificação de possíveis conflitos entre marcas nominativas (BOGAERT *et al.*, 2023). Assim, este estudo oferece uma contribuição relevante para a modernização do sistema brasileiro de registro de marcas, promovendo maior eficiência, segurança jurídica e acessibilidade ao mercado nacional.

Diante desse cenário, esta dissertação investiga a especialização de modelos de

LLM de código aberto para a tarefa de análise de similaridade entre marcas nominativas. Por meio do treinamento supervisionado, os modelos foram adaptados às demandas específicas do contexto de PI, seguindo rigorosamente os critérios estabelecidos pelo INPI. A investigação foi conduzida em duas etapas complementares: a primeira dedicada ao treinamento e avaliação de modelos classificadores, responsáveis pela tarefa de determinar se duas marcas são “*Similares*” ou “*Diferentes*”; e a segunda focada no treinamento de modelos explicadores, que geram justificativas detalhadas para as decisões tomadas, considerando aspectos fonéticos, distintivos, ideológicos e mercadológicos.

A proposta incorpora também uma abordagem interpretável, permitindo fornecer explicações transparentes e fundamentadas sobre a similaridade das marcas analisadas. Além disso, destaca-se a inclusão da análise ideológica, que identifica casos nos quais as marcas, apesar das diferenças gráficas ou textuais, comunicam ideias equivalentes. Esta abordagem representa uma contribuição inédita para a literatura existente, ampliando a compreensão sobre similaridade entre marcas para além dos tradicionais aspectos fonéticos e textuais.

A aplicação prática dessa tecnologia tem potencial para transformar significativamente o processo de análise prévia de marcas, reduzindo o tempo e minimizando custos operacionais. Com a automatização proposta, instituições como o INPI, escritórios especializados em PI e empresas poderão realizar análises mais rápidas, precisas e confiáveis, mitigando riscos de rejeição e conflitos jurídicos. Dessa forma, este estudo apresenta uma contribuição relevante para a modernização do sistema de registro de marcas no Brasil, promovendo maior segurança jurídica, acessibilidade e eficiência no mercado nacional, além de demonstrar o potencial da IA no âmbito da PI.

1.4 Hipótese de Pesquisa

A presente investigação fundamenta-se nas seguintes hipóteses:

- Hipótese 1: LLMs apresentam capacidade para identificar, de maneira precisa e fundamentada, similaridades entre marcas nominativas, considerando os critérios fonéticos, ideológicos, distintivos e mercadológicos, conforme as diretrizes estabelecidas pelo INPI.
- Hipótese 2: LLMs, a partir de engenharia de prompt, conseguem gerar relatórios interpretáveis e detalhados, os quais promovem maior transparência, reduzem a subjetividade e aumentam a eficiência do processo de análise de anterioridade.

1.5 Objetivos

1.5.1 Geral

Propor um método baseado em LLM para a análise de similaridade entre marcas nominativas, composta por dois modelos especializados: um modelo classificador, treinado supervisionadamente para identificar casos de similaridade conforme critérios do INPI, e um modelo explicador, configurado por meio de engenharia de *prompts* para gerar justificativas textuais que fundamentem as decisões, considerando aspectos fonéticos, ideológicos, distintivos e mercadológicos.

1.5.2 Específicos

Os objetivos específicos desta pesquisa de mestrado são:

- Construir um conjunto de dados com casos de litígios registrados pelo INPI no contexto do Brasil;
- Desenvolver um modelo classificador para a avaliação da similaridade entre marcas;
- Desenvolver um modelo explicador, que crie relatórios de colidências considerando aspectos fonéticos, distintivos, ideológicos e mercadológicos;
- Desenvolver um modelo capaz de identificar marcas similares com base nos aspectos ideológicos;
- Avaliar o desempenho dos modelos.

1.6 Organização do Trabalho

Esta dissertação está organizada em sete capítulos, descritos resumidamente a seguir. Destaca-se que um dos capítulos constitui um artigo científico autocontido.

- O **Capítulo 2** apresenta os conceitos fundamentais sobre marcas e PI, marcas, IA e LLMs essenciais para a compreensão do contexto e da relevância deste trabalho;
- O **Capítulo 3** discute estudos relacionados, abordando o estado da arte em análise de similaridade entre marcas e uso de LLMs para análise textual;
- O **Capítulo 4** descreve detalhadamente o processo de criação do conjunto de dados CMAD, cuja metodologia e resultados foram publicados na revista científica **Data**, classificada como A4 na última avaliação Capes;

- O **Capítulo 5** apresenta detalhadamente todas as etapas metodológicas desenvolvidas para atingir os objetivos propostos nesta dissertação;
- O **Capítulo 6** expõe os resultados obtidos, juntamente com a discussão crítica dos achados deste trabalho;
- O **Capítulo 7** traz as considerações finais, destacando as contribuições deste trabalho para as áreas envolvidas, as limitações identificadas, as propostas de trabalhos futuros, bem como as publicações decorrentes deste estudo.

2 Fundamentação Teórica

Este capítulo apresenta, inicialmente, os conceitos fundamentais relacionados à PI, com ênfase em marcas registradas. Em seguida, ele aborda os diferentes tipos de similaridade entre marcas, incluindo similaridade visual, similaridade fonética e textual e similaridade ideológica, destacando sua importância na avaliação de conflitos. Em seguida, são explicados os conceitos de **inteligência artificial** com foco no Processamento de Linguagem Natural (PLN). Por fim, é apresentada uma visão sobre os LLM, explorando sua arquitetura, funcionamento e potencial para a identificação e explicação de similaridades entre marcas nominativas.

2.1 Marcas

Uma MR é um símbolo visual que diferencia os produtos ou serviços de uma empresa daqueles oferecidos pela concorrência, construindo e protegendo sua identidade no mercado e tornando-se um símbolo reconhecido pelos consumidores (INPI, 2023b). Uma marca pode se apresentar na forma nominativa (i.e., um sinal constituído por uma ou mais palavras), figurativa (i.e., um desenho, imagem, figura e/ou símbolo), mista (i.e., uma marca nominativa e uma marca figurativa) e tridimensional (i.e., uma forma plástica distintiva em si mesma) (TRAPPEY; TRAPPEY; LIN, 2020; MACÍAS; CERVIÑO, 2017). Todas elas utilizadas para identificar os produtos ou serviços de uma empresa. A sua definição é altamente consistente em vários institutos de PI em todo o mundo, incluindo o USPTO, o INPI, o CIPO e o WIPO.

Como mostrado na Figura 1, uma marca nominativa, também conhecida como marca verbal, é um sinal distintivo composto por uma ou mais palavras dentro do sistema alfabético romano. Esse tipo de marca pode incluir neologismos, bem como combinações de letras e/ou números romanos e arábicos, desde que não sejam representados de forma estilizada, fantasiosa ou figurativa.

A marca figurativa, também denominada marca emblemática, é caracterizada por ser

TREZENTOS E OITENTA	CORONA	VARIG
Atlântica	IRREVERENTE & CIA	9 E 1/5

Figura 1 – Exemplos de Marcas Nominativas. Fonte: (INPI, 2023b)



Figura 2 – Exemplos de Marcas Figurativas. Fonte: (INPI, 2023b)

composta essencialmente por elementos visuais, como desenhos, imagens, figuras e símbolos, como mostrado na Figura 2. Também se enquadram nessa categoria sinais formados por letras ou algarismos estilizados, de forma fantasiosa, utilizados de maneira isolada ou acompanhados por representações gráficas adicionais. Além disso, são consideradas marcas figurativas aquelas compostas por caracteres oriundos de alfabetos distintos da língua portuguesa, como o hebraico, o cirílico e o árabe, bem como por ideogramas, a exemplo dos utilizados nas línguas japonesa e chinesa (INPI, 2023b).

Uma marca única e distintiva permite a uma empresa ou organização construir e proteger a sua identidade no mercado. Ela torna-se um símbolo que os consumidores associam aos produtos ou serviços específicos da empresa (MACÍAS; CERVIÑO, 2017), considerado um ativo intangível, que garante ao seu proprietário o uso exclusivo sobre ela, assegurando assim que terceiros não beneficiem da eventual reputação que tal símbolo representa, o que fortalece e fideliza os clientes à marca, aumentando as vendas, os preços e, conseqüentemente, melhorando os lucros, uma vez que a maioria dos consumidores tem em conta a reputação da marca quando toma decisões de compra (TRAPPEY; TRAPPEY; LIN, 2020).

Se uma marca não for única, outras empresas podem utilizá-la, causando confusão nos consumidores e prejudicando a reputação da marca original. Esta situação pode dar origem a disputas legais, litígios por infração dos direitos de PI, o que constitui um processo dispendioso e demorado para as partes envolvidas. Logo, a singularidade de uma marca é essencial para proteger os interesses da empresa, evitar a confusão entre os consumidores, uma vez que a semelhança entre a marca requerida e uma marca já registrada é o motivo mais comum para recusar um pedido de registro (SHOWKATRAMANI *et al.*, 2018a).

Vários institutos de PI regionais e nacionais em todo o mundo tratam dos processos de exame e registro de marcas (PROPERTY, ; SHOWKATRAMANI *et al.*, 2018b), como o European Union Intellectual Property Office (EUIPO), USPTO, China National Intellectual Property Administration (CNIPA) e o CIPO. Por exemplo, nos Estados Unidos, quando uma marca solicita o registro, o pedido é enviado a um especialista em PI de marcas para uma análise jurídica exaustiva. Durante a fase de análise, especialistas examinadores podem realizar pesquisas manuais de registros de marcas existentes e de pedidos pendentes anteriores na base de dados do USPTO. Este processo visa determinar o potencial risco de confusão entre a marca solicitada e outras marcas já registradas ou

em processo de registro (USPTO, 2024). O escritório da USPTO, por exemplo, utiliza o sistema de pesquisa interno conhecido como X-search, que consiste em um sistema de recuperação de imagem (SHOWKATRAMANI *et al.*, 2018b).

2.1.1 Aspectos Legais das Marcas no Brasil e Processo de registro

O processo de registro de marcas no Brasil tem início com o depósito do pedido pelo requerente, por meio do sistema eletrônico do INPI, acompanhado do pagamento da respectiva taxa e da indicação da classe de produtos ou serviços com base na Classificação Internacional de Nice. Após o depósito, o pedido é publicado na Revista da Propriedade Industrial (RPI), abrindo-se um prazo de 60 dias para que terceiros possam apresentar oposição, fundamentando a existência de conflito com marcas anteriormente registradas ou requeridas. Caso haja oposição, o requerente será notificado e terá igualmente 60 dias para apresentar sua manifestação. Mesmo que não haja oposição, o pedido segue para a fase de exame técnico formal e substantivo, no qual o examinador avalia se a marca atende aos critérios legais de registrabilidade, incluindo a verificação de conflitos com marcas preexistentes, independentemente da existência de oposição.

Se o examinador identificar conflito com marca anterior, o pedido será indeferido, cabendo ao requerente apresentar recurso administrativo no prazo de 60 dias, o qual será analisado pela instância recursal do INPI. Por outro lado, caso o examinador entenda que não há conflito — mesmo que tenha havido oposição fundamentada —, o pedido poderá ser deferido, e a parte oponente poderá interpor recurso contra o deferimento. Após o deferimento definitivo e o pagamento das taxas correspondentes, o registro é concedido com validade de 10 anos, podendo ser renovado indefinidamente por períodos iguais. É importante destacar que o titular da marca deve acompanhar regularmente a RPI para monitorar pedidos semelhantes que possam comprometer sua exclusividade. Embora o processo possua etapas informatizadas, ele ainda exige análises manuais e técnicas por parte dos examinadores, o que contribui para sua morosidade e demanda acompanhamento jurídico ou especializado constante (INPI, 2023b).

A RPI é um documento oficial publicado semanalmente pelo INPI (INDUSTRIAL, 2023). Essa revista possui diversas seções, sendo uma delas dedicada especificamente aos pedidos de registro de marcas. A seção de marcas exerce um papel fundamental na divulgação de informações relacionadas ao registro de marcas no Brasil, uma vez que publica dados tanto sobre marcas registradas quanto sobre aquelas que foram requeridas. A publicação dessas informações permite que terceiros apresentem oposição ao registro de uma marca, caso entendam que ela possa infringir seus direitos ou gerar conflito com uma marca já existente. Isso garante transparência ao processo de registro e possibilita o acompanhamento e a avaliação dos processos em andamento por parte de terceiros.

2.1.2 Similaridade Fonética e Textual

A presença de reprodução fonética ou de imitação é um dos fatores que frequentemente indicam a existência de um conflito entre duas marcas (LIU *et al.*, 2021). É importante notar que as marcas, incluindo as de apresentação mista, são frequentemente recordadas e mencionadas na sua forma verbal, ou seja, é comum que os consumidores se lembrem e se refiram às marcas por seus nomes, independentemente de elementos gráficos ou visuais que as compõem.

No contexto do INPI, **similaridade fonética** e **similaridade textual** são conceitos distintos, ambos utilizados na análise de colisão entre marcas. A similaridade fonética refere-se à semelhança na pronúncia dos sinais, enquanto a similaridade textual diz respeito à semelhança na grafia ou na forma escrita das marcas. Ambos os aspectos são considerados para avaliar o potencial de confusão entre sinais marcários, mesmo quando não coincidem.

Na avaliação da **similaridade fonética**, consideram-se as semelhanças e diferenças na sequência das sílabas, na entoação das palavras e nos ritmos das expressões presentes nos sinais analisados. Por exemplo, as marcas “*DULCE*” e “*DUCE*” possuem grafias semelhantes, mas são pronunciadas de forma distinta — a primeira com entonação típica do português (/duw./si) e a segunda com uma sonoridade mais estrangeira ou afrancesada (/dy.se/), resultando em percepções auditivas distintas. Por outro lado, as marcas “*NECCHI*” e “*NECHI*” apresentam grafias diferentes, mas são foneticamente idênticas, exemplificando um caso clássico de similaridade fonética. Outro exemplo, as marcas “*VERCOL*” e “*VERKOLL ADD*” apresentam uma pronúncia extremamente próxima, especialmente na sílaba tônica inicial “*VER*” e na terminação sonora “*COL*”/“*KOLL*”, o que pode gerar confusão auditiva entre os consumidores, mesmo com pequenas variações na grafia.

Já a **similaridade textual** refere-se à semelhança visual e ortográfica entre os sinais nominativos. Essa análise considera fatores como número de letras, estrutura silábica, grafia compartilhada e proximidade na sequência de caracteres. Por exemplo, as marcas “*K-MYZZA*” e “*CAMYSA*” apresentam grafias semelhantes, mas podem ser pronunciadas de maneira bastante distinta, demonstrando um caso em que há similaridade textual, mas não fonética.

A similaridade textual pode reforçar ou, em certos casos, contradizer a percepção fonética. Um exemplo que envolve ambas as dimensões é a comparação entre “*ELETRO TECH*” e “*EletriTech*”, cujos elementos gráficos e fonológicos apresentam proximidade significativa — aumentando o risco de confusão por parte do público consumidor. Embora inter-relacionadas, as análises fonética e textual avaliam aspectos distintos. Em conjunto, esses critérios oferecem uma visão mais abrangente da possibilidade de confusão entre marcas, contribuindo para decisões mais criteriosas no processo de registro (INPI, 2023b).

2.1.3 Similaridade Ideológica

Na análise ideológica entre marcas, verifica-se a reprodução ou imitação de um conceito, em que marcas distintas fonética ou graficamente podem evocar ideias semelhantes, gerando potencial confusão no público consumidor. Tal fenômeno ocorre mesmo quando há diferenças significativas na forma de apresentação dos sinais, como ilustrado pelas marcas “PERFECT RUN” e “CORRIDA PERFEITA”. Embora os termos estejam em idiomas diferentes, a semelhança conceitual pode levar o público a associações indevidas, sobretudo considerando-se o nível de familiaridade do consumidor com os idiomas em questão (INPI, 2023b).

Outro exemplo pode ser encontrado na comparação entre as marcas “CIDADE DOS AUTOMÓVEIS” e “CIDADE DOS CARROS”, que apresentam uma notável similaridade ideológica, visto que ambas remetem à ideia de um centro especializado em veículos. Essa semelhança, associada à afinidade mercadológica entre os serviços oferecidos — comércio de automóveis e de peças automotivas — reforça significativamente o risco de confusão. Vale destacar que a análise de afinidade mercadológica não se limita à comparação direta de produtos ou serviços idênticos, abrangendo também bens ou serviços que, apesar de distintos, mantêm relações de complementaridade. Assim, a avaliação do grau de afinidade deve considerar múltiplos fatores, tais como natureza e finalidade dos produtos ou serviços, complementaridade, concorrência, canais de distribuição, público-alvo e origem habitual dos bens ou serviços envolvidos (INPI, 2023b).

2.1.4 Similaridade Visual

O uso de formas geométricas, imagens, cores e/ou combinações de cores semelhantes, pode aumentar o risco de confusão ou associação indevida entre duas marcas. Por isso, é essencial comparar os aspectos gráficos dos sinais ao examinar a possibilidade de conflito. A Figura 3 apresenta um caso em que, apesar das diferenças fonéticas e ideológicas entre os elementos principais das marcas (“GOOD VIBES” e “GOOD TUNES”), é ainda possível identificar semelhanças entre os elementos gráficos secundários. Estas semelhanças gráficas podem potencialmente causar confusão entre elas.

Apesar de ser crucial na análise de marcas figurativas, mistas e tridimensionais, a avaliação das similaridades gráficas também é relevante para sinais nominativos. Nestes casos, a repetição de sequências de letras, o número de palavras e a estrutura das frases e expressões podem, em certas situações, contribuir para a confusão ou associação indevida (INPI, 2023b).



Figura 3 – Exemplo de marcas visualmente semelhantes (INPI, 2023b).

2.2 LLMs

Os Modelos de Linguagem de Grande Escala (LLM, do inglês *Large Language Models*) são redes neurais profundas projetadas para processar e gerar linguagem natural em larga escala, demonstrando um desempenho notável em diversas tarefas linguísticas (MINAEE *et al.*, 2024). Exemplos proeminentes incluem modelos como GPT (BROWN *et al.*, 2020), LLaMA (TOUVRON *et al.*, 2023), Mistral e Gemma, todos fundamentados na arquitetura *Transformer* (VASWANI *et al.*, 2017), que se destaca pela capacidade de capturar dependências linguísticas complexas e aprender representações contextuais profundas das palavras (ZHAO *et al.*, 2024).

A base dos LLM é justamente a arquitetura *Transformer*, cuja principal inovação reside no mecanismo de autoatenção (*self-attention*). Este mecanismo permite ao modelo atribuir pesos distintos a diferentes partes de uma sequência textual, identificando relações entre palavras independentemente da distância que as separa. Ao contrário das redes neurais recorrentes tradicionais, os Transformers conseguem processar sequências textuais de forma paralela, resultando em maior eficiência computacional (HU *et al.*, 2024; ZHAO *et al.*, 2024).

Os LLM são pré-treinados em vastos corpora textuais, como o Common Crawl, utilizando tarefas como predição de palavras e modelagem de linguagem mascarada, em que o modelo aprende a prever uma palavra com base em seu contexto (CHANG *et al.*, 2024). Posteriormente, esses modelos podem ser adaptados para tarefas específicas por meio de *fine-tuning* ou por paradigmas baseados em instruções (*prompting*), sendo este último conhecido como aprendizado em contexto (*in-context learning*). Nesse cenário, o modelo pode aprender a realizar novas tarefas ao receber exemplos diretamente no *prompt*, eliminando a necessidade de treinamento adicional extensivo (MINAEE *et al.*, 2024; CHANG *et al.*, 2024).

A técnica conhecida como ajuste por instruções (*instruction tuning*) permite que os LLM compreendam e sigam comandos complexos, tornando-os altamente eficazes em contextos que demandam interatividade e adaptabilidade elevadas (ZHAO *et al.*,

2024). Essas capacidades emergentes são especialmente úteis em tarefas que exigem compreensão contextual, raciocínio lógico e decomposição de problemas em etapas, tais como sumarização, tradução e respostas a perguntas.

Entretanto, para aplicações críticas, como a classificação em contextos sensíveis, é essencial que os modelos não apenas categorizem os dados com precisão, mas também forneçam explicações interpretáveis para suas decisões. Por exemplo, em análises de saúde mental, modelos como o MentaLLaMA proposto por Yang et al. (YANG *et al.*, 2024b) utilizam LLM para detectar sinais de condições psicológicas a partir de postagens em mídias sociais, fornecendo explicações baseadas em texto. Essas explicações são fundamentais para garantir que decisões automatizadas sejam compreensíveis tanto para especialistas quanto para usuários finais, aumentando a confiança e a transparência dos resultados obtidos (HU *et al.*, 2024).

No entanto, a aplicação dos LLM em classificação e geração de explicações enfrenta desafios consideráveis, particularmente relacionados à qualidade das explicações geradas. Apesar de modelos como o GPT demonstrarem capacidades avançadas de aprendizado em contexto, eles ainda carecem de precisão suficiente em cenários de aprendizado zero-shot ou few-shot. Além disso, a escassez de dados de treinamento de alta qualidade e os elevados custos associados à adaptação dos LLM para domínios específicos representam obstáculos importantes ao avanço dessa tecnologia (YANG *et al.*, 2024b). Uma solução potencial é o *fine-tuning*, que pode alinhar melhor os modelos às especificidades do domínio de aplicação sem comprometer a capacidade de generalização, embora este método exija grandes conjuntos de dados manualmente anotados, o que pode não ser viável em domínios sensíveis.

2.2.1 Fine-tuning de Modelos de Linguagem de Grande Escala (LLMs)

Apesar do excelente desempenho alcançado pelos LLM em diversas tarefas de PLN, frequentemente é necessária uma adaptação adicional para que esses modelos obtenham resultados otimizados em domínios ou aplicações específicas. Este processo de adaptação é denominado *fine-tuning*, e pode ser realizado por meio de três abordagens principais: *full fine-tuning*, ajuste fino com técnicas de aprendizado eficiente (*parameter-efficient fine-tuning*, ou PEFT), e engenharia de prompt (*prompt engineering*) (HU *et al.*, 2024; ZHAO *et al.*, 2024).

O *full fine-tuning* consiste em ajustar todos os parâmetros de um LLM pré-treinado utilizando um conjunto de dados rotulado específico da tarefa alvo. Nesta abordagem, todos os pesos da rede são atualizados, permitindo que o modelo adapte integralmente sua representação às particularidades do novo domínio. Embora este método geralmente alcance excelente desempenho, apresenta como desvantagens principais o alto custo computacional e o risco de *overfitting*, especialmente quando os dados de treinamento são limitados

(ZHAO *et al.*, 2024; YANG *et al.*, 2024b).

Alternativamente, técnicas como o Parameter-Efficient Fine-Tuning (PEFT) têm ganhado popularidade por proporcionarem adaptações eficientes dos modelos com custos reduzidos de treinamento. Diferentemente do *full fine-tuning*, PEFT mantém a maioria dos pesos pré-treinados congelados e ajusta apenas uma fração reduzida dos parâmetros da rede, como em métodos baseados em Low-Rank Adaptation (LoRA), *Prompt Tuning* e *Adapters* (HU *et al.*, 2024). Além disso, abordagens como o *instruction fine-tuning*, uma forma específica de PEFT, ajustam o modelo utilizando instruções explícitas, ensinando-o a interpretar comandos diretos e executar tarefas diversas a partir de poucas interações de treinamento, garantindo grande flexibilidade e robustez ao aprendizado do modelo em múltiplos contextos (CHANG *et al.*, 2024; MINAEE *et al.*, 2024).

Finalmente, a engenharia de prompt representa uma abordagem ainda mais econômica e rápida, na qual não há atualização direta dos pesos do modelo.

Um prompt pode ser definido como uma instrução ou entrada textual fornecida ao modelo com o intuito de direcionar sua geração ou interpretação de conteúdo. Essencialmente, o prompt atua como uma orientação para o modelo, indicando claramente qual tarefa deve ser realizada, seja responder uma pergunta, classificar um texto, resumir um conteúdo ou gerar explicações detalhadas. A qualidade, clareza e especificidade do prompt influenciam diretamente o desempenho do modelo, especialmente em técnicas que não alteram seus parâmetros internos, como o *in-context learning* (LIU *et al.*, 2023). Logo, um prompt bem construído pode induzir comportamentos desejados no modelo, maximizando sua eficácia e garantindo que as respostas sejam relevantes e consistentes com o objetivo proposto.

Para ilustrar o funcionamento da técnica de engenharia de prompt, apresenta-se a seguir um exemplo comparativo. O primeiro prompt é formulado de maneira genérica, solicitando uma explicação ampla sobre o conceito de similaridade entre marcas. Já o segundo prompt, mais elaborado, refina a instrução inicial ao especificar os critérios desejados, exigindo maior profundidade na resposta e fornecendo diretrizes claras ao modelo sobre os aspectos que devem ser abordados.

Prompt Genérico

“Explique o conceito de similaridade entre marcas.”

Prompt Refinado (com Engenharia de Prompt)

“Explique detalhadamente, com exemplos, os critérios fonéticos e ideológicos utilizados pelo INPI para determinar se duas marcas nominativas são consideradas similares ou diferentes.”

Na técnica de engenharia de prompt, o desempenho do modelo é otimizado por

meio da elaboração cuidadosa de prompts que guiam o modelo a gerar respostas adequadas utilizando apenas as capacidades de aprendizado no contexto (*in-context learning*) desenvolvidas durante o pré-treinamento (CHANG *et al.*, 2024). Embora esta técnica seja ágil e exija baixo esforço computacional, sua eficácia pode depender fortemente da qualidade e da formulação precisa dos prompts, e pode não ser suficiente para problemas altamente especializados ou complexos sem o suporte de abordagens mais completas de adaptação (YANG *et al.*, 2024b).

A técnica de *few-shot prompting* é uma abordagem na engenharia de prompts que envolve fornecer ao modelo de linguagem alguns exemplos específicos dentro do próprio prompt para orientá-lo na execução de uma tarefa desejada. Diferentemente do *zero-shot prompting*, onde o modelo recebe apenas a instrução da tarefa sem exemplos adicionais, o *few-shot prompting* utiliza essas demonstrações para condicionar o modelo, melhorando seu desempenho em tarefas mais complexas. Por exemplo, ao ensinar o modelo a usar uma palavra nova em uma frase, podem-se fornecer algumas sentenças de exemplo contendo essa palavra, permitindo que o modelo aprenda o padrão e gere respostas mais precisas.

Few-shot Prompt

Tarefa: Analise se as duas marcas apresentadas são **Similares** ou **Diferentes**, com base em critérios fonéticos e ideológicos, e justifique a resposta.

Exemplo 1

Marcas: *Café do Ponto* vs *Ponto Café*

Resposta: Similares. As marcas possuem os mesmos termos reorganizados, o que mantém o mesmo núcleo semântico. A inversão das palavras não compromete o entendimento fonético nem a ideia central, o que pode gerar confusão no consumidor.

Exemplo 2

Marcas: *TechNova* vs *Nova Tecnologia*

Resposta: Similares. Apesar da variação morfológica, ambas as marcas evocam a mesma ideia central de inovação tecnológica, e compartilham radicais fonéticos comuns.

Agora analise o seguinte par de marcas:

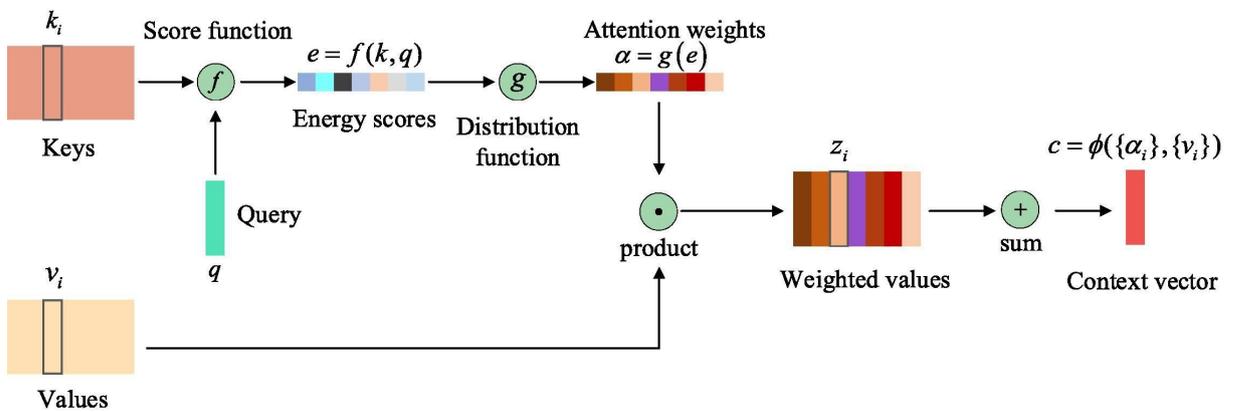
Marcas: *Mais Fit* vs *Fit+*

2.2.2 Arquitetura dos LLMs

A base estrutural dos LLMs) é a arquitetura *Transformer*, introduzida por Vaswani *et al.* (2017). Essa abordagem inovadora superou significativamente métodos anteriores, como as Redes Neurais Recorrentes (RNN) e as Long Short-Term Memory (LSTM) (do português, Redes de Memória de Longo Prazo), ao permitir processamento paralelo eficiente das sequências textuais. O *Transformer* fundamenta-se no mecanismo de *self-attention*, que atribui pesos distintos às palavras de uma sequência, capturando eficientemente o contexto linguístico, mesmo em relações de longo alcance (HU *et al.*, 2024).

Internamente, a arquitetura *Transformer* é composta por múltiplas camadas de mecanismos de atenção e redes *feed-forward*, estabelecendo uma estrutura robusta para a modelagem da linguagem natural (VASWANI *et al.*, 2017). O mecanismo de *self-attention* permite interações diretas entre todas as palavras de uma frase, assegurando uma representação mais profunda e contextualizada dos elementos textuais (GALASSI; LIPPI; TORRONI, 2021). Complementarmente, o uso da *multi-head attention* possibilita que o modelo aprenda simultaneamente múltiplas perspectivas contextuais sobre a mesma informação, aprimorando assim sua capacidade de generalização e compreensão semântica (KAMATH; LIU; WHITAKER, 2019; NIU; ZHONG; YU, 2021). A Figura 4 apresenta um esquema da estrutura básica do modelo de atenção dentro da arquitetura *Transformer*.

Figura 4 – Estrutura do mecanismo de *attention* utilizado na arquitetura *Transformer*.
Fonte: Niu, Zhong e Yu (2021)



Além das inovações arquiteturais, o desempenho excepcional dos LLM é também atribuído ao seu treinamento extensivo em grandes conjuntos de dados textuais. Técnicas como a Modelagem de Linguagem Mascarada (MLM) utilizada pelo BERT (KOROTEEV, 2021), e a predição autoregressiva adotada pelos modelos GPT (RADFORD *et al.*, 2018; BROWN *et al.*, 2020), são exemplos notáveis de estratégias de pré-treinamento que possibilitam a captura eficaz de regularidades e nuances linguísticas. Após essa etapa, os modelos são frequentemente ajustados (*fine-tuning*) para realizar tarefas específicas com alta precisão e relevância contextual (ZHAO *et al.*, 2024).

Atualmente, diferentes variações do *Transformer* são aplicadas nos LLM, destacando-se três principais arquiteturas: modelos exclusivamente codificadores (*encoder-only*), como o BERT; exclusivamente decodificadores (*decoder-only*), como GPT-3 e LLaMA; e codificador-decodificador, representados pelo modelo T5 (*Text-To-Text Transfer Transformer*) (RAFFEL *et al.*, 2020; YANG *et al.*, 2024a). Cada variação arquitetural impacta diretamente as capacidades e a adequação dos modelos para diferentes tipos de tarefas no PLN.

O aumento contínuo no tamanho e complexidade desses modelos tem resultado em avanços notáveis, porém trouxe desafios significativos relacionados ao consumo ener-

gético, viés algorítmico e alta demanda computacional. O treinamento desses modelos frequentemente requer o uso intensivo de Graphics Processing Unit (GPU) e Tensor Processing Unit (TPU), destacando a importância de aprimorar a eficiência e sustentabilidade computacional nessa área (ZHAO *et al.*, 2023).

Recentemente, houve ainda o desenvolvimento e incorporação de abordagens multimodais nos LLMs, como exemplificado pelos modelos GPT-4V e Gemini. Esses modelos integram a capacidade de processamento textual com imagens, áudios e vídeos, ampliando significativamente sua aplicabilidade e potencializando a criação de sistemas mais integrados e versáteis (ZHANG; BOULOS, 2023). Dessa forma, a pesquisa na arquitetura dos LLM permanece ativa, orientando-se tanto para avanços tecnológicos quanto para a resolução de desafios práticos e éticos associados ao uso desses modelos em contextos reais (ZHAO *et al.*, 2024).

3 Trabalhos Relacionados

Este capítulo apresenta os trabalhos relacionados à tarefa de análise de similaridade entre marcas. A literatura atual é composta majoritariamente por métodos convencionais, como Redes Neurais Convolucionais (CNNs), algoritmos fonéticos e modelos baseados em *word embeddings*, aplicados à comparação textual ou fonética entre marcas. Essas abordagens focam, em geral, na extração de características específicas — ortográficas, fonéticas ou visuais — com o objetivo de avaliar o grau de semelhança entre marcas. Embora LLMs estejam sendo amplamente utilizados para tarefas de similaridade textual, até o momento, não foram identificados trabalhos que explorem seu uso diretamente na tarefa de identificação de similaridade entre marcas.

3.1 Soluções baseadas em LLMs para similaridade textual

Os LLMs emergiram como ferramentas poderosas para análise de similaridade textual devido à sua capacidade avançada de compreender contextos linguísticos e modelar relações semânticas complexas. Essa capacidade é adquirida por meio do treinamento em larga escala em grandes volumes de texto, nos quais os modelos aprendem representações estatísticas profundas das relações entre palavras e expressões, permitindo-lhes capturar nuances semânticas com alto grau de precisão (CHANG *et al.*, 2024).

Modelos generativos baseados na arquitetura Transformer destacam-se especialmente por utilizar mecanismos autoregressivos em suas camadas *decoder*, permitindo-lhes gerar textos coerentes e contextualizados. Em contrapartida, modelos não-generativos, como o *Bidirectional Encoder Representations from Transformers* (BERT), também alcançam excelente desempenho na análise semântica através de suas camadas *encoder*, projetadas especificamente para produzir representações contextuais densas, embora não tenham capacidade direta de geração textual. Ambos os tipos de modelos, portanto, são capazes de realizar comparações semânticas eficazes, ainda que utilizem abordagens distintas para alcançar essa capacidade (MINAEE *et al.*, 2024).

Uma abordagem não-generativa é o Sentence-BERT (SBERT), proposto por (REIMERS; GUREVYCH, 2019), que consiste em uma modificação da arquitetura BERT, utilizando redes siamesas e estruturas de tripletos para gerar *embeddings* de sentenças semanticamente significativos. Esses *embeddings* permitem uma comparação eficiente entre sentenças por meio da métrica de similaridade cosseno, reduzindo drasticamente o tempo de processamento em tarefas de busca de similaridade semântica, sem comprometer a precisão dos resultados.

De forma complementar, modelos generativos também foram explorados recentemente na criação de datasets sintéticos para melhorar o desempenho na tarefa de similaridade textual. O Sim-GPT, desenvolvido por (WANG *et al.*, 2023c), aborda especificamente a escassez de conjuntos de dados rotulados de alta qualidade para tarefas de Similaridade Textual Semântica (STS). A estratégia adotada utiliza o GPT-4 para gerar um volume substancial de dados anotados com pontuações de similaridade textual. Com base nesses dados sintéticos, um modelo especializado em STS é treinado utilizando BERT ou RoBERTa como base. O Sim-GPT alcança desempenho superior em benchmarks de STS, superando modelos anteriores como o SimCSE e o PromCSE.

Finalmente, destacando aplicações relacionadas que utilizam abordagens semelhantes baseadas em LLMs, (HU *et al.*, 2024) buscam aprimorar tarefas de classificação textual, especificamente a verificação de autoria de produção textual, por meio de técnicas eficientes de *fine-tuning*, como o método LoRA. Embora focado em verificação de autoria, esse estudo demonstra claramente a eficácia do ajuste fino de LLMs para melhorar não apenas a acurácia da classificação, mas também a qualidade das explicações geradas pelos modelos, fortalecendo a justificativa para utilizar técnicas semelhantes em análises detalhadas de similaridade textual.

3.2 Soluções para Similaridade Fonética e Textual de Marcas

Estudos têm explorado a aplicação de métodos convencionais de Aprendizado de Máquina (AM) para a detecção de similaridade fonética e textual entre marcas, com destaque para abordagens baseadas em CNNs e modelos de representação de texto.

Ko *et al.* (2018) propõem um método baseado em CNNs para avaliação da similaridade fonética de marcas. A técnica envolve a transcrição fonética das marcas multilíngues, seguida pela segmentação dessas transcrições em *bigrams* (pares consecutivos de caracteres). Esses *bigrams* são convertidos em representações fonéticas bidimensionais, servindo como entrada para o treinamento das redes convolucionais. O método alcançou aproximadamente 92% de precisão em um conjunto de dados desenvolvido especificamente para este estudo, contendo marcas semelhantes e distintas.

Por outro lado, Trappey, Trappey e Lin (2020) abordam a avaliação da similaridade textual e fonética por meio de técnicas mais tradicionais. Para similaridade textual, empregam o modelo Word2Vec para gerar *word embeddings*, utilizando a similaridade cosseno como métrica comparativa. Já a similaridade fonética é avaliada com algoritmos clássicos, como *Soundex* e *Metaphone*, combinados ao cálculo de similaridade pelo algoritmo de *Levenshtein*. Esses métodos foram testados independentemente em um conjunto de 166 casos de litígios extraídos do banco de dados jurídico *Westlaw*. No entanto, comparações diretas com outras abordagens não foram possíveis devido à exclusividade e especificidade

do conjunto de dados utilizado.

Em uma abordagem mais ampla, Liu *et al.* (2021) desenvolvem um modelo baseado em CNNs capaz de detectar similaridades semânticas, fonéticas e visuais entre marcas compostas por caracteres chineses. Nesse método, cada marca é representada por meio de *embeddings* especializados que capturam aspectos semânticos, fonéticos, visuais e estruturais das palavras. Essas representações são concatenadas e utilizadas como entrada em uma rede neural siamesa, treinada para identificar marcas similares. Avaliado em um extenso conjunto de dados contendo 130.000 pares de marcas similares e 242.000 pares distintos, o modelo atingiu uma precisão média de 87,1%.

A Tabela 1 apresenta uma comparação entre os principais trabalhos relacionados à tarefa de avaliação de similaridade entre marcas, considerando os aspectos metodológicos, os tipos de dados utilizados, os critérios de similaridade adotados e a presença de modelos explicativos. Observa-se que os estudos anteriores concentram-se predominantemente na similaridade textual ou fonética, com abordagens baseadas em CNNs, algoritmos fonéticos tradicionais ou *word embeddings*. Em contraste, o trabalho proposto nesta dissertação adota uma abordagem utilizando LLM de código aberto, treinados com dados reais do INPI, para avaliar não apenas aspectos fonéticos e textuais, mas também ideológicos e mercadológicos. Além disso, diferencia-se por incorporar um modelo explicador capaz de gerar justificativas detalhadas para as decisões.

3.3 Contribuições desta Pesquisa

Esta pesquisa diferencia-se significativamente das demais ao propor um *framework* baseado em LLM especializado na análise de similaridade entre marcas nominativas, destacando-se por integrar dois modelos complementares: um modelo classificador e um modelo explicador. As principais contribuições e inovações deste trabalho são destacadas a seguir:

- **Modelo Classificador:** Desenvolvimento de um modelo classificador treinado de forma supervisionada para avaliar a similaridade entre marcas nominativas. O modelo é capaz de identificar com elevada precisão os casos de colisão, tendo aprendido, a partir de um conjunto de dados reais rotulado por especialistas do INPI, as nuances fonéticas, ideológicas, distintivas e mercadológicas que orientam as decisões técnicas no processo de exame.
- **Modelo Explicador Interpretável:** Utilização de engenharia de *prompts* para configuração de um modelo explicador, que gera relatórios textuais detalhados com justificativas fundamentadas para cada decisão tomada pelo modelo classificador. Esses relatórios explicativos proporcionam maior transparência, reduzindo subjetivi-

Tabela 1 – Comparação entre trabalhos relacionados na avaliação de similaridade entre marcas.

Trabalho	Técnica Utilizada	Aspectos Considerados	Tipo de Dados	Modelo Explicativo	Precisão Reportada
(KO <i>et al.</i> , 2018)	CNN com transcrição fonética e <i>bigrams</i>	Similaridade fonética	Marcas multilíngues transcritas foneticamente	Não	92%
(TRAPPEY; TRAPPEY; LIN, 2020)	Word2Vec (textual), Soundex/-Metaphone + Levenshtein (fonética)	Similaridade textual e fonética	Casos jurídicos do Westlaw	Não	Não especificado
(LIU <i>et al.</i> , 2021)	CNN + <i>embeddings</i> semânticos, fonéticos e visuais em rede siamesa	Similaridade semântica, fonética e visual	Caracteres chineses	Não	87,1%
Este trabalho (Proposto)	LLMs open-source com <i>fine-tuning</i> supervisionado e engenharia de <i>prompt</i>	Similaridade fonética, textual, ideológica e mercadológica	Marcas nominativas brasileiras (casos reais do INPI)	Sim (modelo explicador baseado em LLM)	≈99%

dades e facilitando a compreensão das decisões, considerando os critérios fonéticos, distintivos, ideológicos e mercadológicos estabelecidos pelo INPI.

- **Análise Ideológica como Inovação:** Introdução da análise ideológica como um aspecto inovador na literatura, permitindo aos modelos identificar similaridade entre marcas com base em ideias transmitidas, mesmo quando há diferenças fonéticas ou gráficas significativas.
- **Utilização de LLMs:** Aplicação de LLMs, que representam o estado-da-arte em inteligência artificial, potencializando a capacidade de avaliação precisa e fundamentada da similaridade entre marcas.

Além disso, esta pesquisa traz contribuições adicionais relevantes:

- **Conjunto de Dados Exclusivo:** Construção e utilização de um conjunto de dados

único e especializado, composto por casos reais de colisões registrados pelo INPI, junto com relatórios revisados por especialistas em PI.

- **Flexibilidade Linguística e Escalabilidade:** O *framework* proposto pode ser aplicado para análise de marcas em diversos contextos linguísticos e em diferentes escalas de aplicação. Embora especializado para o contexto brasileiro, conforme diretrizes do INPI, o modelo possui potencial para adaptação em outros contextos jurisdicionais e linguísticos.

Essas inovações proporcionam um avanço substancial na eficiência, transparência e objetividade do processo de verificação de marcas. Ao reduzir significativamente o esforço manual e os custos operacionais, o *framework* também mitiga riscos associados a litígios e aumenta a agilidade na análise de anterioridade. Este estudo, portanto, contribui decisivamente para a literatura sobre análise de similaridade entre marcas.

4 Conflicting Marks Archive Dataset

Este capítulo tem o objetivo de apresentar o CMAD (sigla para *Conflicting Marks Archive Dataset*), um conjunto de dados voltado para casos de litígio em experimentos de similaridade entre marcas. Os casos selecionados foram aqueles em que os pedidos de registro foram indeferidos pelo INPI, com o intuito de evitar conflitos entre marcas e reduzir a possibilidade de litígios. Este conjunto de dados foi utilizado para desenvolver o método proposto neste trabalho. Acreditamos que o CMAD será um instrumento fundamental para impulsionar, tanto esta, como outras pesquisas e avanços na área de PI.

4.1 Conjunto de Dados Relacionados

Na literatura, os termos “marca” e “logotipo” são distintos e possuem significados específicos. Marca refere-se a um sinal distintivo, como um nome, símbolo ou desenho, utilizado para identificar os produtos ou serviços de uma empresa e distingui-los daqueles de outras empresas, estando legalmente protegido junto aos órgãos competentes para impedir o uso não autorizado por terceiros (OMPI, ; INDUSTRIAL, 2023; CIPO, 1976). O termo logotipo, por sua vez, refere-se a um elemento gráfico ou símbolo que representa uma empresa ou organização e que, geralmente, compõe a identidade visual da marca, mas que ainda não está legalmente protegido (KESIDIS; KARATZAS, 2014).

As tarefas conhecidas na literatura como “Detecção de Logotipo”, “Recuperação de Imagem” e “Similaridade entre Marcas” são distintas entre si, embora possam se relacionar dependendo de sua aplicação. A detecção de logotipo refere-se ao processo de identificar a presença ou localização de logotipos específicos em uma imagem ou vídeo (BAO *et al.*, 2016; PSYLLOS; ANAGNOSTOPOULOS; KAYAFAS, 2010). A recuperação de imagem refere-se à tarefa de encontrar imagens similares ou relevantes em um conjunto de dados com base em uma imagem de consulta fornecida pelo usuário (SMEULDERS *et al.*, 2000). A ideia é recuperar imagens que compartilham características visuais ou semânticas com a imagem de referência. Já a similaridade entre marcas refere-se ao grau em que duas marcas são semelhantes visual, fonética ou conceitualmente (INDUSTRIAL, 2023).

Na literatura, existem diferentes conjuntos de dados originalmente desenvolvidos para as tarefas de detecção de logotipo e recuperação de imagem, sendo que alguns estudos os adaptaram para aplicação na tarefa de similaridade entre marcas (ALSHOWAISH; AL-OHALI; AL-NAFJAN, 2022; VESNIN; LEVSHUN; CHECHULIN, 2023; LAN *et al.*, 2017). A seguir, descrevemos os conjuntos de dados criados para essas tarefas.

- *BelgaLogos* (JOLY; BUISSON, 2009; JOLY; BUISSON, 2023) foi criado para a tarefa

de detecção de logotipos. Ele contém 10.000 imagens anotadas manualmente de 26 logotipos. Cada imagem é rotulada para cada logotipo, com 1 indicando a presença do logotipo na imagem e 0 caso contrário. As imagens do conjunto podem conter um ou vários logotipos, ou nenhum. O conjunto de teste contém o nome da marca presente na imagem, o nome do arquivo e as coordenadas dos pixels que delimitam o logotipo presente na imagem;

- *FlickrLogos-32* (ROMBERG *et al.*, 2011; ROMBERG *et al.*, 2023) contém imagens de 32 logotipos distintos e seus respectivos rótulos. Ele foi criado e dividido em três conjuntos distintos, denominados P1, P2 e P3. O primeiro conjunto (P1) destina-se ao treinamento de algoritmos de AM/Aprendizado Profundo (AP) e possui 10 imagens por classe, contendo logotipos sob diferentes perspectivas. Os conjuntos P2 e P3, respectivamente para validação e teste (ou consulta), contêm 30 imagens por classe, nas quais há pelo menos uma instância de logotipo;
- *Logo-2K+* (WANG *et al.*, 2020; WANG *et al.*, 2023a) foi desenvolvido para a tarefa de detecção de logotipos e possui um total de 167.140 imagens. As imagens pertencem a marcas divididas em 10 categorias (por exemplo, alimentos, vestuário, instituições, acessórios), e subdivididas em 2.341 subcategorias, representando cada marca;
- *LogoDet-3K* (WANG *et al.*, 2022; WANG *et al.*, 2023b) foi criado para a detecção de logotipos e contém 3.000 categorias de logotipos, com cerca de 200.000 objetos de logotipo anotados manualmente e 158.652 imagens. As imagens dos logotipos são divididas em 9 categorias (ou seja, alimentos, roupas, utilidades, eletrônicos, transporte, lazer, esportes, saúde e outros), e subdivididas em 3.000 subcategorias;
- *LOGO-Net* (HOI *et al.*, 2015; HOI, 2015) é um grande conjunto de dados de imagens para detecção de logotipos, incluindo dois conjuntos com um total de 81.874 imagens: o conjunto “logos-18” possui um total de 16.043 objetos de logotipo em 8.460 imagens; e o conjunto “logos-160” contém um total de 130.608 objetos de logotipo em 73.414 imagens. Eles foram criados por meio de *web scraping* (ou seja, coleta automatizada na Internet (KUMAR; BHATIA; RATTAN, 2017)) em sites de shoppings virtuais, sendo cada imagem posteriormente anotada manualmente para delimitar a região do logotipo;
- *METU* (TURSUN; AKER; KALKAN, 2017; TURSUN; AKER; KALKAN, 2023) é um conjunto de dados desenvolvido para a tarefa de recuperação de imagens e possui 923.343 imagens de diferentes tipos: logotipos compostos apenas por texto, apenas por figuras, e por figuras com texto. Possui dois conjuntos principais: o conjunto de consulta (*query set*) e o conjunto de teste. O conjunto de consulta contém 417 imagens de marcas rotuladas manualmente e agrupadas por similaridade em 35 classes.

A Tabela 2 apresenta uma análise comparativa dos conjuntos de dados relacionados, incluindo suas respectivas tarefas, número de amostras e tipos de dados. Os conjuntos de dados analisados não são especificamente voltados para a tarefa de similaridade entre marcas, mas sim para as tarefas de detecção de logotipos e recuperação de imagens. O número de amostras nesses conjuntos varia de menos de 10 mil até quase 1 milhão. Além disso, os conjuntos relacionados são compostos exclusivamente por imagens, sem incluir quaisquer dados tabulares, descritores ou metadados.

Tabela 2 – Análise comparativa entre conjuntos de dados relacionados e o conjunto de dados desenvolvido neste trabalho.

Dataset	Tarefa	Número de Amostras	Tipo de dados
<i>BelgaLogos</i> (JOLY; BUISSON, 2009; JOLY; BUISSON, 2023)	Detecção de Logotipo	10,000	Imagens
<i>FlickrLogos-32</i> (ROMBERG <i>et al.</i> , 2011; ROMBERG <i>et al.</i> , 2023)	Detecção de Logotipo	8,240	Imagens
<i>Logo-2K+</i> (WANG <i>et al.</i> , 2020; WANG <i>et al.</i> , 2023a)	Detecção de Logotipo	167,140	Imagens
<i>LogoDet-3K</i> (WANG <i>et al.</i> , 2022; WANG <i>et al.</i> , 2023b)	Detecção de Logotipo	158,652	Imagens
<i>LOGO-NET</i> (HOI <i>et al.</i> , 2015; HOI, 2015)	Detecção de Logotipo	81,874	Imagens
<i>METU</i> (TURSUN; AKER; KALKAN, 2017; TURSUN; AKER; KALKAN, 2023)	Recuperação de Imagens	930,328	Imagens
CMAD	Similaridade de Marcas	Número de pares = 18,355	Imagens e Dados Tabulares

Diferentemente dos conjuntos de dados relacionados, o CMAD foi criado especificamente para a tarefa de similaridade entre marcas. Ele foi produzido com base em marcas que foram requeridas para registro no INPI, mas que tiveram seus pedidos indeferidos em razão de algum tipo de conflito. Esses indeferimentos foram cuidadosamente analisados por especialistas do INPI. Assim, a similaridade entre marcas no CMAD representa conflitos rotulados de forma confiável.

Além disso, para o melhor do nosso conhecimento, o CMAD é o primeiro conjunto de dados a reunir casos de marcas colidentes com base em três tipos de similaridade (ou seja, nominativa, ideológica e visual), incluindo infrações marcárias entre marcas nominativas, figurativas e mistas. É importante destacar que marcas nominativas (isto é, compostas por texto) podem representar a maioria dos registros em escritórios de PI (LIU *et al.*, 2021).

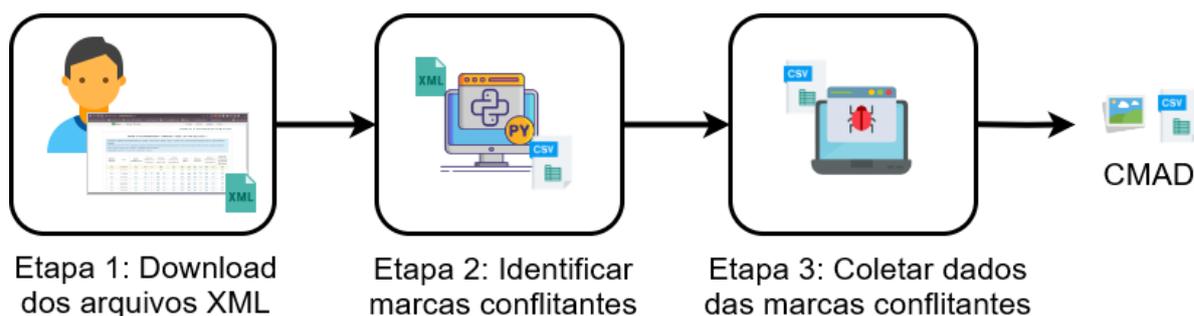


Figura 5 – Etapas da metodologia.

Adicionalmente, diferentemente dos conjuntos BelgaLogos (JOLY; BUISSON, 2009; JOLY; BUISSON, 2023), FlickrLogos-32 (ROMBERG *et al.*, 2011; ROMBERG *et al.*, 2023) e LogoDet-3K (WANG *et al.*, 2022; WANG *et al.*, 2023b), o CMAD é composto por imagens de marcas e dados tabulares, organizados de forma a facilitar o acesso a informações relacionadas às marcas colidentes. O CMAD oferece uma base sólida para o desenvolvimento e teste de ferramentas voltadas à análise de similaridade entre marcas, contribuindo, assim, para o avanço das práticas e soluções computacionais no campo da PI. Portanto, ao escolher o CMAD como conjunto de dados para tarefas de similaridade entre marcas, pesquisadores e profissionais da área jurídica podem se beneficiar não apenas da diversidade e confiabilidade dos casos reais de litígio entre marcas, mas também de sua estrutura organizacional e tabular.

4.2 Métodos

A metodologia deste estudo envolve a análise e extração de dados a partir das revistas publicadas pelo INPI. O fluxograma representado na Figura 5 descreve as etapas da metodologia.

Na Etapa 1, realizamos manualmente o download das revistas em formato Extensible Markup Language (XML). Esse arquivo é utilizado como entrada para o algoritmo que realiza a análise e extração dos pedidos. Na revista, há uma seção que descreve os pedidos indeferidos e a respectiva motivação, conforme ilustrado na Figura 6.

829220267	<p>Indeferimento do pedido Titular: EPI - EMPREENDIMENTO PATRIMONIAL INDUSTRIAL S/A [BR/PR] Procurador: Manoel Paixao do Nascimento NCL(9): 21 Especificação: ESPONJAS DE LIMPEZA PARA USO DOMÉSTICO, ESCOVAS PARA LOUÇAS, ESCOVAS PARA CALÇADOS, ESCOVAS PARA ESFREGAR, VASSOURAS. (DA CLASSE 21)</p> <p>Detalhes do despacho: A marca reproduz ou imita os seguintes registros de terceiros, sendo, portanto, irregistrável de acordo com o inciso XIX do Art. 124 da LPI: Processo 828906564 (BRILLMAX).</p>
------------------	---

Figura 6 – Exemplo de indeferimento.

Na Etapa 2, desenvolvemos um algoritmo escrito na linguagem de programação Python que recebe o arquivo XML (Figura 7) como entrada e identifica conteúdos relacionados aos pedidos de marca (por exemplo, número do processo, texto complementar), utilizando expressões regulares. Esse algoritmo gera um arquivo Comma-Separated Values (CSV) contendo as marcas colidentes.

```

<processo numero="927779820">
  <despachos>
    <despacho codigo="IPAS024" nome="Indeferimento do pedido">
      <texto-complementar>A marca reproduz ou imita os seguintes registros de terceiros, sendo,
      portanto, irregistrável de acordo com o inciso XIX do Art. 124 da LPI: Processo 918753252 (YUPY KIDS),
      Processo 921945086 (YUUUP! BABY) e Processo 830036970 (IUPIII!! BEBÊ).
      Art. 124 - Não são registráveis como marca: XIX - reprodução ou imitação, no todo ou em parte, ainda que com acréscimo,
      de marca alheia registrada, para distinguir ou certificar produto ou serviço idêntico, semelhante ou afim,
      suscetível de causar confusão ou associação com marca alheia;</texto-complementar>
    </despacho>
  </despachos>
  <titulares>
    <titular nome-razao-social="IZAURA MARIA VICTORINO DA SILVA" pais="BR" uf="PR"/>
  </titulares>
  <lista-classe-nice>
    <classe-nice codigo="35">
      <especificacao>Comércio (através de qualquer meio) de artigos do vestuário;Comércio (através de qualquer meio) de fitas e laços; </esp
      <status>Deferida</status>
    </classe-nice>
  </lista-classe-nice>
  <procurador>Danielle Juliá Lopes Brites</procurador>
</processo>

```

Figura 7 – Exemplo de um arquivo XML contendo um pedido de registro de marca.

Na Etapa 3, cada pedido indeferido é analisado. No texto complementar do pedido, identificamos e extraímos as marcas colidentes e, em seguida, iniciamos o processo de coleta de dados no site do INPI por meio de um *crawler*. Para esse fim, desenvolvemos outro algoritmo em Python utilizando a biblioteca Selenium (CONTRIBUTORS, 2023). O *crawler* coleta inicialmente informações sobre o pedido indeferido (por exemplo, número do processo, nome, forma de apresentação, natureza, classificação de Nice), conforme ilustrado na Figura 8. Quando a forma de apresentação da marca é figurativa ou mista, o *crawler* também coleta a imagem da marca. Em seguida, o algoritmo coleta as mesmas informações da(s) marca(s) colidente(s). Caso haja mais de uma marca colidente, os dados são coletados para todas elas. Ao final, é gerado um arquivo CSV contendo a marca indeferida correspondente a cada marca colidente.

4.3 Descrição dos Dados

O CMAD contém 18.355 amostras de conflitos entre marcas, nas quais uma marca requerida para registro entrou em conflito com uma marca já registrada. O conjunto de dados possui um diretório com imagens em formato Portable Network Graphic (PNG) e um arquivo CSV contendo dados tabulares relacionados a cada amostra. No arquivo CSV, cada amostra está estruturada em pares e inclui 8 colunas para cada marca (ou seja, a marca indeferida e a marca já registrada no INPI), além de um texto complementar e da referência à revista. Para as colunas que se referem à marca registrada, seus cabeçalhos são seguidos da sigla “TM” (isto é, *TradeMark*) e, para as colunas referentes à marca

BRASIL | Acesso à informação | Participe | Serviços | Legislação | Canais

Instituto Nacional da Propriedade Industrial
Ministério da Economia

Consulta à Base de Dados do INPI

[Início | Ajuda?]

» Consultar por: No.Processo | Marca | Titular | Cód. Figura] 1/0

Marca

Nº do Processo: **829220267**

Marca: BRILMAX

Situação: Aguardando apresentação e exame de recurso contra o indeferimento

Apresentação: Mista

Natureza: De Produto



Classificação de Produtos / Serviços

Classe de Nice	Situação da Classe	Especificação
NCL(9) 21	Vide Situação do Processo	ESPONJAS DE LIMPEZA PARA USO DOMÉSTICO, ESCOVAS PARA LOUÇAS,...

Figura 8 – Exemplo de dados coletados de uma marca indeferida.

indeferida, os cabeçalhos são seguidos da sigla “RM” (isto é, *Rejected Mark*). As duas colunas finais, texto complementar e revista, referem-se, respectivamente, à justificativa apresentada para a oposição ao registro da marca e ao número da revista. A Tabela 3 apresenta as colunas do conjunto de dados. Ressalta-se que o arquivo CSV foi criado em língua portuguesa do Brasil, com dados provenientes do INPI.

A forma mista de apresentação corresponde à maior parte dos registros no conjunto de dados. As marcas indeferidas (Figura 9a) somam 14.391 (78,4%) na forma mista e 3.964 (21,6%) na forma nominativa. Já as marcas registradas (Figura 9b) totalizam 13.021 (70,9%) na forma mista e 5.334 (29,1%) na forma nominativa.

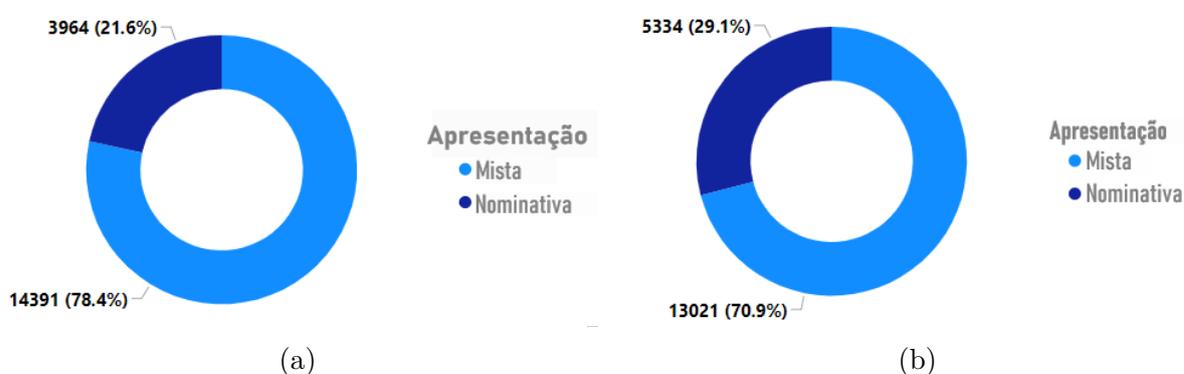


Figura 9 – Distribuição das amostras quanto à forma de apresentação: (a) pedidos de marcas indeferidas; e (b) marcas registradas.

A Figura 10 ilustra três amostras de marcas colidentes, com os pares de imagens

Tabela 3 – Descrição dos dados do arquivo CSV do CMAD.

Nome do Campo	Descrição	Tipo
<i>Process number</i>	Número do processo atribuído ao pedido de registro, utilizado para identificar de forma única a marca no conjunto de dados, bem como para acessar o caminho da imagem, quando houver (isto é, não há imagens para marcas nominativas).	Numérico
<i>Name</i>	Nome da marca.	Texto
<i>Status</i>	Status da marca (por exemplo, marca indeferida, registrada, aguardando análise).	Texto
<i>Presentation</i>	Tipo de apresentação da marca (por exemplo, Nominativa, Figurativa e Mista).	Texto
<i>Nature</i>	Natureza da marca (por exemplo, produtos, serviços).	Texto
<i>Nice classification</i>	Classificação internacional de produtos e serviços, adotada globalmente e gerenciada pela WIPO. É utilizada para categorizar as marcas de acordo com sua área de atuação (ORGANIZATION, 2024a; PROPERTY, 2024).	Texto
<i>Vienna classification</i>	Classificação internacional gerenciada pela WIPO para categorizar elementos gráficos em marcas figurativas, mistas e tridimensionais. Auxilia na descrição e especificação de elementos visuais durante o registro de marcas, evitando conflitos e proporcionando clareza (ORGANIZATION, 2024b).	Texto
<i>Application date</i>	Data em que a marca foi requerida.	Data
<i>Complementary</i>	Texto que descreve as razões pelas quais o pedido foi indeferido.	Texto
<i>Magazine</i>	Número da publicação da revista.	Texto

das marcas e, respectivamente, suas entradas no arquivo CSV.

4.4 Discussão

4.4.1 Aplicações do CMAD

O CMAD é útil em diversas áreas. Na indústria, tem especial valor para o desenvolvimento e teste de softwares voltados à avaliação de similaridade entre marcas. Especificamente, o CMAD pode auxiliar no desenvolvimento e aprimoramento de sistemas projetados para comparar e identificar semelhanças entre marcas.

Na academia, o uso do CMAD é altamente benéfico em estudos voltados à análise de similaridade marcária, seja para fins de validação ou de treinamento de algoritmos de AM / AP. Diversos estudos (LIU *et al.*, 2021; LAN *et al.*, 2018; TRAPPEY; TRAPPEY; LIN, 2019; TRAPPEY; TRAPPEY; SHIH, 2021) se destacam por desenvolver modelos



(a)

Process number RM	Name RM	Status RM	Presentation	Process number TM	Name TM	Status TM	Presentation TM
924058803	My Bank\$	Aguardando apresentação	Mista	918356849	MEUBANK	Registro de marca em vigor	Mista
925224022	AUTOBOX LUBRIFICANTES	Aguardando apresentação	Mista	908946481	AUTOBOX	Registro de marca em vigor	Mista
927878143	SOCORRO AUTO CHICO	Aguardando apresentação	Mista	912622962	AUTO SOCORRO Silva & Silva	Registro de marca em vigor	Mista

(b)

Figura 10 – Três amostras de marcas colidentes no CMAD: (a) três pares de imagens de marcas; e (b) respectivas entradas no arquivo CSV.

de DL que utilizam pares de dados similares durante o treinamento, com o objetivo de avaliar a similaridade entre marcas. Esses trabalhos exemplificam o potencial do CMAD em contribuir efetivamente para o desenvolvimento de modelos de AM / AP. Ao adotar o CMAD como fonte de dados, pesquisadores têm a oportunidade de explorar uma ampla variedade de casos de conflito entre marcas.

Embora o CMAD seja focado na similaridade entre marcas, sua aplicabilidade pode ser mais ampla. Por exemplo, pode ser utilizado na área de recuperação de imagens. Diversos estudos (VELMURUGAN; BABOO, 2011; CAO *et al.*, 2021; TURSUN *et al.*, 2022; LAN *et al.*, 2017) utilizam conjuntos de dados de marcas para implementar algoritmos avançados de extração de características, com o objetivo de melhorar a eficiência e a

precisão na identificação de marcas. O conjunto de imagens do CMAD se apresenta como uma opção viável para o desenvolvimento de algoritmos de extração de características voltados à tarefa de recuperação de imagens. Dessa forma, o nosso conjunto de dados pode oferecer diversidade e representatividade para o treinamento e validação desses algoritmos.

4.4.2 Considerações Éticas

O CMAD inclui apenas características das marcas, como nome, classificação e imagem, não contendo qualquer informação pessoal sobre seus titulares. Esses dados já se encontram disponíveis publicamente na plataforma online do INPI. Portanto, não há impedimentos éticos ou legais quanto ao uso dessas informações para fins de pesquisa e análise. A utilização do CMAD assegura o cumprimento das normas éticas e legais ao excluir dados pessoais sensíveis. O uso do conjunto de dados está alinhado com o Plano de Dados Abertos do INPI (INDUSTRIAL, 2022; INDUSTRIAL, 2024), que estabelece diretrizes para a implementação e promoção da abertura dos dados do instituto.

4.4.3 Forças e Limitações

Ao concentrar-se exclusivamente em casos de conflito entre marcas no Brasil, o CMAD oferece uma perspectiva única e valiosa para análises no contexto nacional. Isso significa que os dados refletem nuances específicas do mercado brasileiro, tornando o CMAD particularmente relevante para estudos e aplicações voltados ao país. Aspectos legais e culturais locais desempenham um papel crucial na formação das marcas. Por exemplo, a forma como as marcas são percebidas e interpretadas pode ser influenciada por fatores culturais, como idioma, simbolismo e normas sociais.

Além disso, a legislação brasileira de PI, que rege o registro e a proteção das marcas, pode ter implicações significativas para a existência e resolução de conflitos entre marcas. Portanto, o uso dos dados do CMAD deve considerar esses aspectos. A compreensão desses elementos pode auxiliar na identificação de tendências, padrões e insights mais precisos e contextualizados a um cenário nacional.

5 Materiais e Métodos

Neste capítulo, são descritos os métodos empregados para desenvolver e validar os LLMs. Para alcançar o objetivo deste trabalho, propõe-se um método que utiliza dois modelos LLMs: um modelo classificador e um modelo explicador. O modelo classificador é responsável por analisar pares de marcas e determinar se são similares ou distintas, fornecendo uma resposta objetiva: “sim” ou “não”. Já o modelo explicador tem a função de gerar relatórios detalhados que justificam as decisões do modelo classificador com base nos critérios como fonética, ideologia, distintividade e afinidade mercadológica. Essa abordagem visa não apenas melhorar a acurácia da detecção de similaridade, mas também garantir maior transparência e interpretabilidade no processo de análise, tornando os resultados mais compreensíveis para especialistas em PI. A Figura 11 apresenta uma visão geral do método implementado nesta pesquisa.

O método proposto foi estruturado em quatro etapas principais. Inicialmente, foi realizada a engenharia de prompts voltada aos modelos classificador e explicador, elaborados a partir de casos reais de conflitos entre marcas registrados pelo INPI, utilizando o conjunto de dados CMAD (REIS *et al.*, 2024) (ver Capítulo 4). Em seguida, nove LLMs foram submetidos a fine-tuning para a tarefa de classificação, visando identificar aqueles com melhor desempenho na avaliação da similaridade entre marcas.

Na terceira etapa, os prompts elaborados para a geração de explicações detalhadas foram submetidos ao GPT-4o, considerado o mais avançado disponível no momento da pesquisa. As respostas geradas por este modelo foram validadas por especialistas em PI por meio de um sistema web desenvolvido especificamente para esse fim, permitindo ajustes e reformulação dos prompts. Por fim, os três melhores modelos da etapa de classificação foram submetidos à técnica de engenharia de prompt com abordagem *few-shot*, utilizando como referência os relatórios validados anteriormente pelos especialistas, garantindo assim

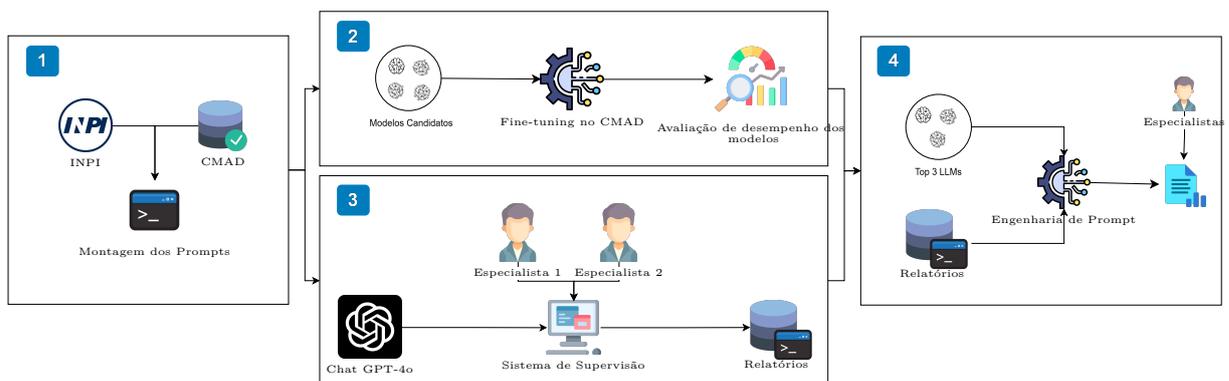


Figura 11 – Visão Geral da Metodologia.

uma geração precisa e alinhada às diretrizes do INPI. Cada uma dessas etapas será apresentada em detalhes nas seções a seguir.

5.1 Elaboração dos prompts

Na primeira etapa, a partir das marcas conflitantes reunidas no conjunto CMAD, foram elaborados dois formatos de prompt: um destinado à classificação binária, no qual o modelo deve indicar se duas marcas são consideradas similares, e outro voltado para geração do relatório de similaridade, fornecendo através do prompt diretrizes para a geração do relatório sobre os critérios que justificam a relação entre as marcas avaliadas. O modelo classificador recebe duas marcas como entrada no prompt e tem a função de determinar se são similares, fornecendo como saída apenas “Sim” ou “Não”. Esse processo busca garantir uma classificação objetiva e direta, sem qualquer justificativa adicional. Dessa forma, o primeiro prompt foi estruturado da seguinte maneira:

Prompt para modelo classificador

Aqui estão duas marcas: [MARCA 1] [MARCA 2]

Essas duas marcas são similares? Responda apenas com “Sim” ou “Não”.

SOLUÇÃO

A resposta correta é:

Por sua vez, o modelo explicador recebe como entrada no prompt as duas marcas e informações sobre o setor de atuação, incluindo a classe nice e sua descrição. A classe nice da marca refere-se à classe da Classificação Internacional de Nice, que especifica os produtos ou serviços aos quais a marca se aplica, definindo seu setor de atuação no mercado. Sua função é gerar uma justificativa detalhada sobre os fatores que levaram ou não à identificação de conflito entre as marcas. Para elaborar esse prompt, consideramos as diretrizes estabelecidas no Manual de Marcas, redigido e publicado pelo INPI (INPI, 2023b). Esse manual fornece os critérios oficiais utilizados na avaliação da similaridade entre marcas, garantindo que o modelo opere de acordo com as normas e diretrizes adotadas pelos especialistas em PI. Há quatro aspectos fundamentais na análise da similaridade entre marcas: a análise fonética, ideológica, de distintividade e de afinidade mercadológica. Logo, o seguinte prompt, solicitando a análise de marcas similares, foi estruturado da seguinte maneira:

Prompt para modelo explicador

Realize uma análise da similaridade entre as marcas nominativas '[MARCA 1]' ([DESCRIÇÃO NICE DA MARCA 1]) e '[MARCA 2]' ([DESCRIÇÃO NICE DA MARCA 2]) e explique por que essas marcas são semelhantes e podem causar risco de confusão ou associação indevida entre os nomes.

Considere os seguintes aspectos:

1. Aspecto fonético: Avalie a pronúncia dos nomes levando em conta a entonação, ritmo e sequência das sílabas.
2. Aspecto ideológico: Verifique se os nomes das marcas evocam ideias semelhantes, levando em consideração a classe Nice associada, mesmo que suas pronúncias sejam diferentes.
3. Distintividade: Determine se os nomes possuem elementos dominantes semelhantes.
4. Afinidade mercadológica: Analise a proximidade entre os produtos ou serviços associados às marcas, com base nas classes Nice correspondentes.

Para avaliar duas marcas diferentes, o prompt é:

Prompt para modelo explicador

Realize uma análise da similaridade entre as marcas nominativas '[MARCA 1]' ([DESCRIÇÃO CLASSE NICE DA MARCA 1]) e '[MARCA 2]' ([DESCRIÇÃO NICE DA MARCA 2]) e explique porque essas marcas são diferentes e não causam risco de confusão ou associação indevida entre os nomes.

Considere os seguintes aspectos:

1. Aspecto fonético: Avalie a pronúncia dos nomes levando em conta a entonação, ritmo e sequência das sílabas.
2. Aspecto ideológico: Verifique se os nomes das marcas evocam ideias semelhantes, levando em consideração a classe Nice associada, mesmo que suas pronúncias sejam diferentes.
3. Distintividade: Determine se os nomes possuem elementos dominantes semelhantes.
4. Afinidade mercadológica: Analise a proximidade entre os produtos ou serviços associados às marcas, com base nas classes Nice correspondentes.

A principal diferença entre os dois prompts está na orientação inicial da análise, ou seja, na afirmação que o modelo explicador deverá justificar: se as marcas são similares ou diferentes. Como o modelo classificador já determina a similaridade ou não entre as marcas, o modelo explicador é instruído a elaborar um relatório que justifique a decisão com base nos quatro critérios estabelecidos.

5.2 Fine-tuning e de Avaliação dos modelos de LLMs classificadores

Na etapa 2, selecionamos nove LLMs *open-source*, optando exclusivamente por suas versões *instruct*, em vez da versão *base*. Os modelos base são treinados em vastos conjuntos de dados de texto, abrangendo desde sites até livros e artigos, para aprender os padrões e estruturas da linguagem. Embora sejam proficientes em gerar texto coerente, eles não são especificamente projetados para seguir instruções diretas, o que pode resultar em respostas que nem sempre atendem às expectativas do usuário. Esses modelos servem como fundação para desenvolvimentos posteriores, podendo ser refinados para aplicações específicas. Já os modelos *instruct* passam por um treinamento adicional supervisionado, onde aprendem a seguir comandos específicos e a fornecer respostas mais alinhadas às solicitações humanas. Esse processo envolve o uso de pares de instruções e respostas durante o treinamento, melhorando a capacidade do modelo de interpretar e responder a tarefas de maneira precisa e consistente. Logo, optou-se por essa versão visto que os modelos são mais eficientes em seguir instruções complexas, executar tarefas com múltiplas etapas e responder com um estilo ou formato específico (CUCONASU *et al.*, 2024).

O uso dos modelos se deu através da biblioteca *Unsloth* (AI, 2024), que disponibiliza modelos quantizados e implementa a técnica LoRA, permitindo otimizar o uso de recursos computacionais durante o treinamento. Cada um desses modelos foi submetido a um processo de fine-tuning, sendo testado quanto à sua capacidade de classificar corretamente marcas conflitantes do CMAD com “sim” ou “não”, garantindo que o modelo escolhido apresentasse o melhor desempenho para essa tarefa. Os experimentos ocorreram em um computador com duas placas de vídeo RTX 4090 com 24GB de RAM cada, sendo utilizada apenas uma delas durante o experimento; memória RAM de 128 GB e processador Intel Core i9 de 13^a geração e 2TB de SSD.

Para otimizar o uso de recursos computacionais, adotamos modelos de linguagem otimizados por meio da técnica de quantização *bnb-4bit* (*Bitsandbytes 4-bit quantization*), um método que reduz significativamente o consumo de memória sem comprometer o desempenho dos LLMs. Logo, tornou-se viável realizar o treinamento de modelos que tradicionalmente exigem alto poder computacional em máquinas com menor capacidade de processamento. A sigla *bnb* refere-se à biblioteca *Bitsandbytes*, uma implementação de quantização e otimização para modelos de AP (DETTMERS, 2025), permitindo a execução desses modelos em hardware com capacidade limitada ao reduzir o uso de memória VRAM. A quantização *4-bit* consiste em armazenar os pesos dos modelos com uma precisão de 4 bits, em contraste com os tradicionais 16 ou 32 bits utilizados em modelos de maior precisão. Essa técnica impacta diretamente no tamanho do modelo, diminuindo a quantidade de memória necessária para sua execução e tornando viável sua utilização em dispositivos com restrições computacionais, como GPUs de menor capacidade.

A Tabela 4 detalha as especificações de cada modelo. A coluna Modelo identifica o

nome de cada modelo, enquanto Origem indica a instituição desenvolvedora. Em Tamanho, tem-se a quantidade aproximada de parâmetros em bilhões (B).

Tabela 4 – Modelos de LLMs selecionados para avaliação

Modelo	Empresa	Tamanho
Qwen2-7B	Alibaba	7B
Llama-3-8B	Meta	8B
Mistral-7B	Mistral	7B
Gemma-2-9B	Google	9B
Mistral-Nemo	Mistral	7B
Phi-3-Medium	Microsoft	7B
Meta-Llama-3.1-8B	Meta	8B
Phi-3.5-Mini	Microsoft	2B
Llama-3.2-1B	Meta	1B

Na etapa de treinamento, utilizamos a técnica Efficient Finetuning of Quantized LLMs (QLoRA) para otimizar o *fine-tuning* dos LLMs selecionados. A QLoRA é uma técnica baseada na abordagem LoRA, que além de reduzir significativamente o número de parâmetros treináveis, aplica a quantização dos pesos do modelo original para valores de baixa precisão (4 bits). Dessa forma, o método permite uma adaptação extremamente eficiente e acessível de modelos pré-treinados, reduzindo consideravelmente o consumo de memória e o custo computacional durante o treinamento (HU *et al.*, 2021).

A técnica LoRA original realiza ajustes no modelo pré-treinado por meio da adaptação de baixa classificação, isto é, adapta os pesos originais adicionando matrizes de baixa dimensão, mantendo os parâmetros originais inalterados e treinando apenas essas matrizes adicionais. A QLoRA, por sua vez, aprimora essa abordagem ao aplicar quantização dos pesos originais, representando-os com menor precisão numérica — tipicamente em 4 bits — o que reduz substancialmente o uso de memória. Essa técnica possibilita o ajuste fino de modelos de grande escala mesmo em ambientes com recursos computacionais limitados, mantendo desempenho equivalente ao ajuste fino tradicional em 16 bits, conforme evidenciado pelos resultados no benchmark MMLU, em que o modelo QLoRA obteve desempenho médio de 53,1%, comparado a 53,0% do modelo em BFloat16 (DETTMERS *et al.*, 2023).

Neste estudo, utilizamos os valores padrão da biblioteca *unsloth* para a configuração da QLoRA, garantindo um ajuste fino otimizado. O raio (r) foi mantido em 16, restringindo a adaptação de baixa classificação a um subespaço reduzido, enquanto *lora_alpha* permaneceu em 16, regulando a escala das atualizações para preservar a estabilidade do aprendizado. A configuração incluiu um *batch size* de 64, sem acumulação de gradientes, totalizando 100 passos de treinamento.

Para otimizar o desempenho do modelo e evitar sobreajuste, implementamos um mecanismo personalizado de *Early Stopping*. Esse método monitora a perda do treinamento e interrompe a execução caso o erro atinja um valor menor do que zero. O treinamento foi

realizado utilizando a classe *SFTTrainer* da biblioteca *unsloth*. A configuração utilizada incluiu 32 amostras por lote (*batch size*), acumulação de gradientes em 1 etapa, aquecimento (*warmup*) de 10 passos e um agendador de taxa de aprendizado baseado em cosseno. Além disso, foi empregado o otimizador *AdamW* com acumuladores em 8 bits, visando minimizar o consumo de memória.

Para realizar o *fine-tuning* e a avaliação dos modelos, foi utilizado o conjunto de dados CMAD, composto por 18.355 exemplos de pares de marcas consideradas conflitantes. Visando garantir o balanceamento das classes na tarefa de classificação binária, foram gerados aleatoriamente 18.355 exemplos adicionais contendo pares de marcas que não apresentam conflito.

Para isso, a partir dos nomes das marcas registradas e das marcas conflitantes, desenvolveu-se um algoritmo para selecionar aleatoriamente marcas não relacionadas entre si. Esse algoritmo percorre a base de marcas registradas válidas e, de forma controlada e aleatória, combina nomes distintos que não compartilham vínculo de conflito conhecido, montando pares que representam casos de ausência de similaridade.

Dessa forma, o conjunto final totalizou 36.710 instâncias, sendo metade de pares conflitantes e metade de pares não conflitantes. Os dados foram estruturados com base nos prompts de classificação descritos na Seção 5.1, assegurando que cada exemplo estivesse formatado adequadamente para permitir a detecção precisa da similaridade entre as marcas.

Para viabilizar o treinamento, foi necessário converter os rótulos dos conflitos para um formato numérico. Implementamos um processo de mapeamento que atribui o valor 1 para casos em que as marcas foram classificadas como “Similar”, e 0 para aquelas consideradas “Diferente”. Com os dados devidamente organizados, realizamos uma amostragem estratificada, distribuindo 80% dos exemplos para o conjunto de treinamento (29.368 instâncias) e 20% para o conjunto de teste (7.342 instâncias), como apresentado na Figura 12

Para avaliar os modelos quantitativamente, a saída do modelo foi obtida na forma de *logits*, correspondentes às classes 0 (Diferente) e 1 (Similar). Para converter essas saídas em probabilidades, aplica-se a função softmax, garantindo que os valores fossem interpretáveis como distribuições de probabilidade. A classe final foi determinada pelo índice da maior probabilidade. Para avaliar o desempenho dos modelos, as métricas utilizadas foram acurácia, precisão, *recall*, *F1-Score* e Area Under the Curve (AUC). Além disso, também realizamos a análise de curva Receiver Operating Characteristic (ROC).

A Tabela 5 apresenta a matriz de confusão, base para calcular as métricas de desempenho. Ela organiza as previsões do modelo em quatro categorias: os verdadeiros positivos (TP), que representam casos positivos corretamente identificados; os falsos

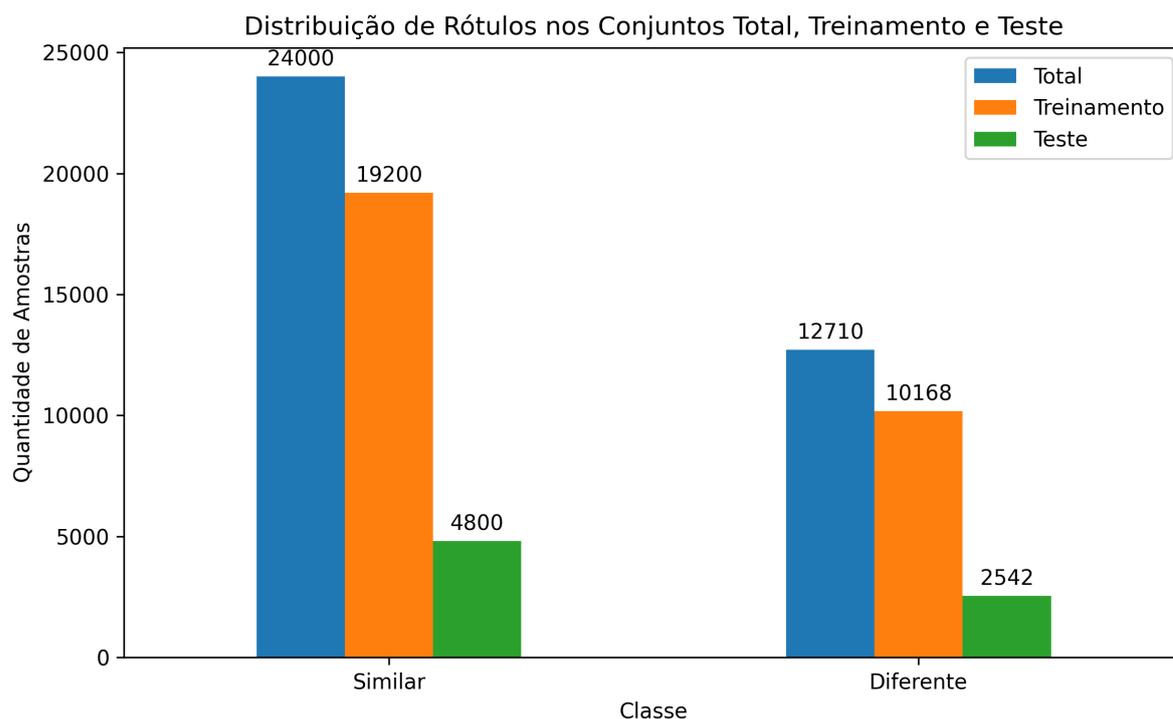


Figura 12 – Divisão da base de dados.

positivos (FP), casos negativos classificados incorretamente como positivos; os falsos negativos (FN), casos positivos classificados incorretamente como negativos; e os verdadeiros negativos (TN), casos negativos corretamente classificados.

Tabela 5 – Matriz de confusão.

		Valores reais		Total
		Positivo	Negativo	
Valores previstos	Positivo	<i>Verdadeiro Positivo</i>	<i>Falso Positivo</i>	$VP + FP$
	Negativo	<i>Falso Negativo</i>	<i>Verdadeiro Negativo</i>	$FN + VN$
Total		$VP + FN$	$FP + VN$	

A acurácia é a proporção geral de acertos obtidos pelo modelo em relação ao total de casos avaliados, definida conforme a Equação 5.1. A precisão indica a proporção de casos classificados como positivos que foram corretamente identificados pelo modelo. Em outras palavras, mede quantos dos resultados positivos previstos são realmente positivos, conforme a Equação 5.2. O *recall*, também conhecido como *Sensitivity* (sensibilidade), representa a proporção de casos positivos reais que o modelo foi capaz de identificar corretamente. Essa métrica está representada pela Equação 5.3. O *F1-score* ou *F-measure* é a média harmônica entre precisão e *recall*. É uma métrica muito utilizada para equilibrar o desempenho do modelo, especialmente quando as classes são desbalanceadas, sendo definida pela Equação 5.4.

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$

$$Precisão = \frac{TP}{TP + FP} \quad (5.2)$$

$$Recall = \frac{TP}{TP + FN} = Sensitivity \quad (5.3)$$

$$F - score = \frac{2 * (Precisão * Recall)}{Precisão + Recall} \quad (5.4)$$

A curva ROC é uma ferramenta gráfica amplamente utilizada para avaliar o desempenho de modelos de classificação binária. Essa curva ilustra a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos, em diferentes limiares de decisão. A métrica AUC representa numericamente a área sob a curva ROC, variando entre 0 e 1, em que valores próximos a 1 indicam excelente desempenho discriminativo do modelo, enquanto valores próximos a 0,5 sugerem desempenho próximo ao aleatório.

5.3 Avaliação por Especialista

Na terceira etapa, os prompts desenvolvidos para a tarefa de classificação na primeira fase foram submetidos ao GPT-4o, um dos mais avançados LLMs disponíveis no momento da pesquisa. As respostas geradas foram posteriormente avaliadas por especialistas em PI. Para facilitar esse processo, foi desenvolvido um sistema web no qual os especialistas puderam revisar as respostas, indicando se concordavam ou não com as justificativas fornecidas.

Cada especialista recebeu credenciais exclusivas para acessar a plataforma de avaliação. Participaram do processo um doutor em PI e um profissional que atua em avaliação de marcas em escritórios especializados. Para cada instrução, foi gerado um relatório de colidência pelo GPT-4o, que posteriormente foi submetido à avaliação dos especialistas, que indicaram se concordavam ou não com o conteúdo do relatório. Em caso de discordância, eles realizaram modificações no texto para ajustá-lo conforme suas exigências e garantir maior precisão na análise. A Figura 13 mostra uma captura de tela da ferramenta utilizada nesta etapa.



Figura 13 – Tela de avaliação do relatório.

5.4 Engenharia de Prompt com Abordagem Few-Shot: Geração e Avaliação de Relatórios por Especialistas

Na quarta etapa, os três LLMs com melhor desempenho na etapa 2 foram utilizados para gerar relatórios de colidência, os quais foram posteriormente avaliados por especialistas em PI. Nesta fase, aplicamos a técnica de engenharia de prompt, onde o modelo foi refinado diretamente por meio do prompt de entrada, utilizando a abordagem de *few-shot prompting*. Para isso, dois relatórios previamente avaliados por especialistas na etapa 3 foram incorporados ao prompt, servindo como exemplos para orientar o modelo na geração dos relatórios subsequentes. Um exemplo do prompt pode ser consultado no Apêndice A

Para cada modelo, foi solicitado que gerasse um total de 10 relatórios, sendo 5 justificando por qual motivo duas marcas são similares e 5 explicando a razão delas serem diferentes, resultando em 30 relatórios ao todo. Esses relatórios foram então submetidos à avaliação de 4 especialistas em PI, que analisaram a qualidade das explicações fornecidas. Para cada relatório gerado, os especialistas realizaram a avaliação utilizando uma escala, variando de 0 a 5, em que 0 representa a nota mínima (total desacordo) e 5 corresponde à nota máxima (total concordância). Além disso, foi disponibilizado um campo adicional para cada relatório para que os especialistas pudessem registrar comentários sobre os relatórios avaliados, permitindo observações, críticas ou sugestões de aprimoramento.

6 Resultados e Discussão

Neste capítulo, apresentamos e analisamos quantitativamente e qualitativamente os resultados obtidos nas duas etapas do método proposto. Primeiramente, são apresentados os resultados de desempenho dos nove LLMs na tarefa de classificação binária de similaridade entre marcas. Em seguida, são apresentados os resultados da avaliação dos relatórios gerados pelos três melhores modelos na geração de relatórios explicativos, com base nos critérios estabelecidos. Por fim, discutimos criticamente os resultados alcançados, destacando suas implicações práticas da abordagem proposta no contexto da análise de marcas.

6.1 Modelos Classificadores

Os modelos apresentaram um desempenho bastante variável, destacando a importância da escolha do modelo adequado para a tarefa de detecção de similaridade entre marcas. A Figura 14a ilustra as matrizes de confusão dos três modelos que obtiveram melhor desempenho: Mistral-Nemo, Qwen2-7B e Llama-3-8b. Observa-se que, em todos esses casos, a maior parte das amostras foi classificada corretamente, evidenciando a robustez dos modelos na distinção entre marcas “Similares” e “Diferentes”.

Já a Figura 14b e a Figura 14c apresenta as matrizes de confusão dos demais seis modelos avaliados. Nota-se que alguns modelos, como o Llama-3.1-8b e o Llama-3.2-1b, também alcançaram resultados satisfatórios, com baixo número de erros. Em contraste, modelos como Mistral-7B, Phi-3-medium-4k e Phi-3.5-mini falharam completamente em classificar corretamente a classe “Similar”, atribuindo todas as amostras à classe “Diferente”.

A Tabela 6 apresenta o desempenho dos modelos para cada métrica. Alguns modelos, como Llama-3-8B, Qwen2-7B e Mistral-Nemo, alcançaram altos valores de precisão e F1-score, enquanto outros, como Mistral-7B, Phi-3-Medium e Phi-3.5-Mini, tiveram desempenho extremamente baixo. Os modelos Mistral-Nemo, Llama-3-8B e Qwen2-7B apresentaram a melhor acurácia, o que indica a proporção de previsões corretas feitas pelos modelos em relação ao total de previsões. Mistral-Nemo e Qwen2-7b apresentam também a melhor precisão, indicando a quantidade de verdadeiros positivos que o modelo realmente acertou. Ou seja, contabiliza a quantidade de vezes que o modelo realmente acertou ao dizer que duas marcas eram realmente similar.

Em relação ao recall, que mede a quantidade de vezes que o modelo acertou ao identificar todos os positivos reais, ou seja, das marcas realmente similares, quantas o modelo conseguiu encontrar, o Gemma-2-9B teve o melhor desempenho (0,9937), no

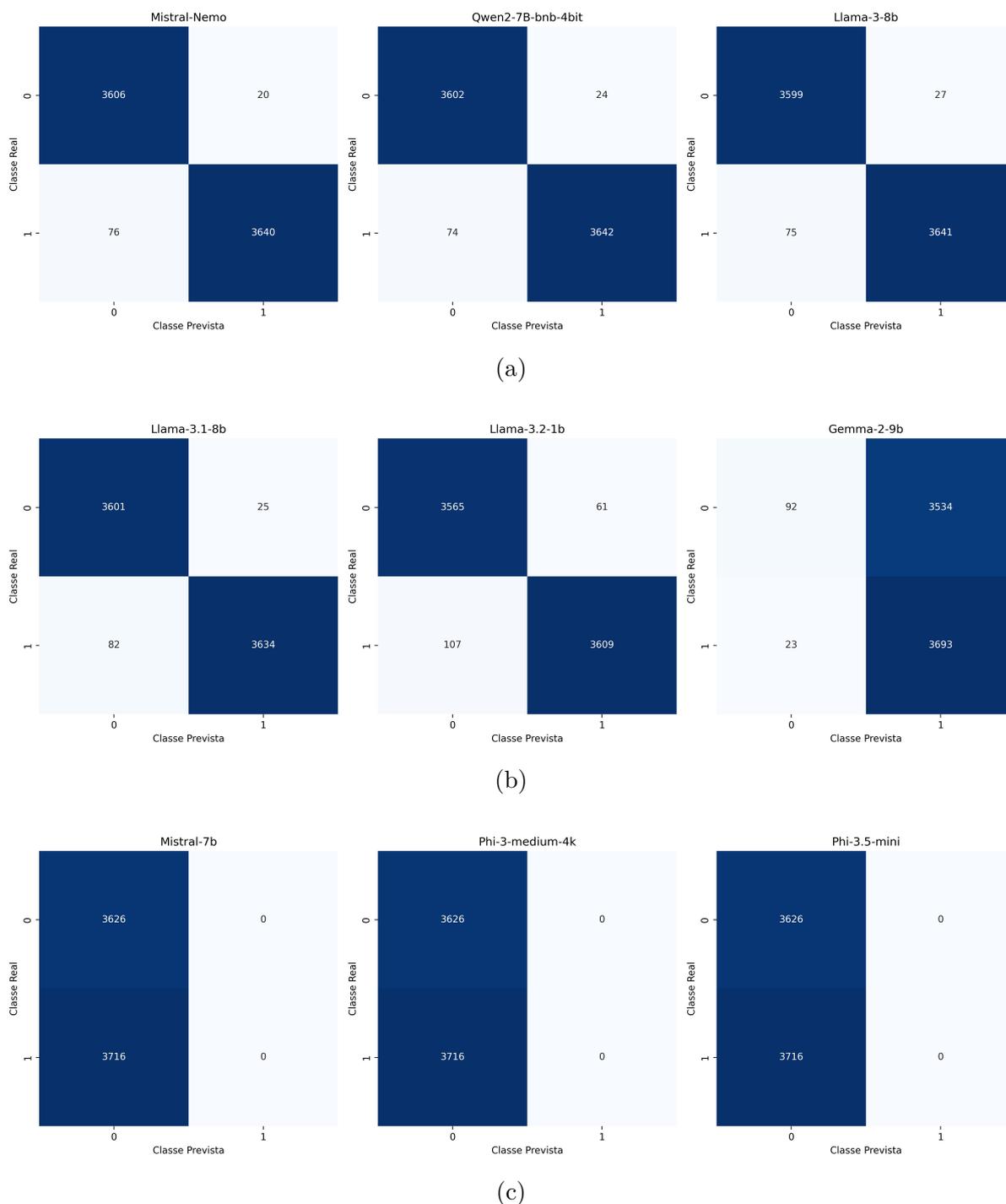


Figura 14 – Matrizes de confusão dos modelos avaliados.

entanto, a com a sua acurácia e recall medindo 0,5155 e 0,5109 respectivamente, ou seja, quase metade da amostra, mostra que o modelo classificou praticamente todas as amostras como similares.

Os modelos Llama-3-8B e Qwen2-7B se destacaram, com acurácia de 0,9860 e 0,9867 respectivamente e F1-score superior a 0,986, indicando que ambos foram eficazes na distinção entre marcas similares e diferentes. Já o modelo Mistral-Nemo acurácia e F1-score ligeiramente superior, 0,9869 e 0,987, respectivamente, sendo o modelo mais

Tabela 6 – Métricas de desempenho dos modelos avaliados, ordenadas pelo F1-Score.

Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Mistral-Nemo	0,9869	0,9946	0,9795	0,9870	0,9983
Qwen2-7B	0,9867	0,9936	0,9800	0,9867	0,9978
Llama-3-8B	0,9860	0,9925	0,9797	0,9861	0,9981
Llama-3.1-8B	0,9853	0,9930	0,9779	0,9854	0,9982
Llama-3.2-1B	0,9772	0,9835	0,9713	0,9773	0,9947
Gemma-2-9B	0,5155	0,5109	0,9937	0,6749	0,7467
Phi-3.5-Mini	0,4939	0,0000	0,0000	0,0000	0,5298
Mistral-7B	0,4939	0,0000	0,0000	0,0000	0,3841
Phi-3-Medium	0,4939	0,0000	0,0000	0,0000	0,3519

consistente na tarefa.

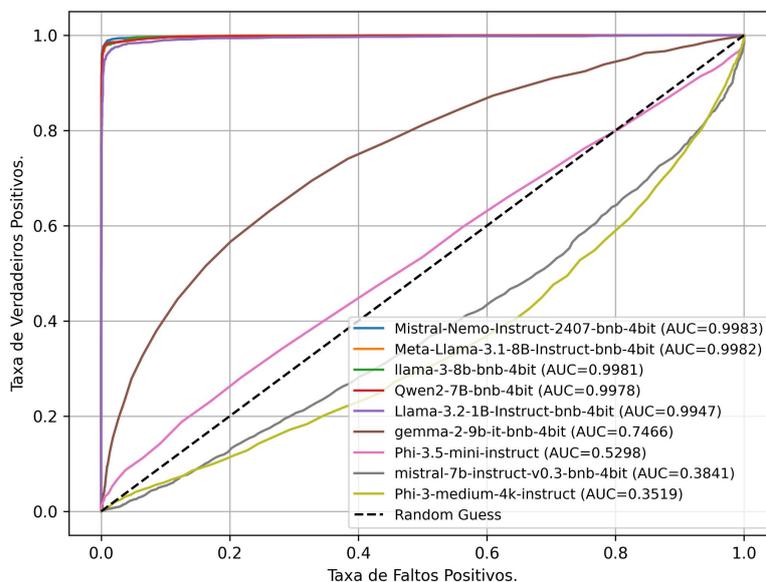
Por outro lado, modelos como Mistral-7B, Phi-3-Medium e Phi-3.5-Mini apresentaram desempenho extremamente baixo, com F1-score, precisão e recall igual a 0,0. Isso ocorre porque esses modelos classificaram todas as instâncias em uma única classe, resultando em uma matriz de confusão completamente assimétrica, onde nenhum caso foi corretamente identificado como “Similar”. O modelo Gemma-2-9B, embora tenha apresentado alta taxa de recall (0,9937), teve acurácia de apenas 0,5155 e F1-score de 0,6749.

A métrica AUC fornece uma avaliação mais abrangente da capacidade discriminativa dos modelos ao longo de diferentes limiares de decisão. Um valor de AUC próximo de 1.0 indica que o modelo tem uma alta capacidade de distinguir entre classes, enquanto um valor próximo de 0.5 sugere um desempenho semelhante ao de uma classificação aleatória. O Mistral-Nemo obteve o melhor resultado, com AUC de 0.9983, seguido de perto pelo Llama-3.1 (0,9982) e pelo Llama-3-8B (0,9981). O modelo Qwen2-7B também apresentou um desempenho robusto, com um AUC de 0,9978, demonstrando uma capacidade de discriminação próxima à dos modelos com maiores parâmetros. Por outro lado, modelos com baixa acurácia e F1-score, como Gemma-2-9B, Phi-3.5-Mini, Mistral-7B e Phi-3-Medium, apresentaram valores de AUC baixos, com atenção aos modelos Mistral-7B e Phi-3-Medium, com AUC inferior a 0,4.

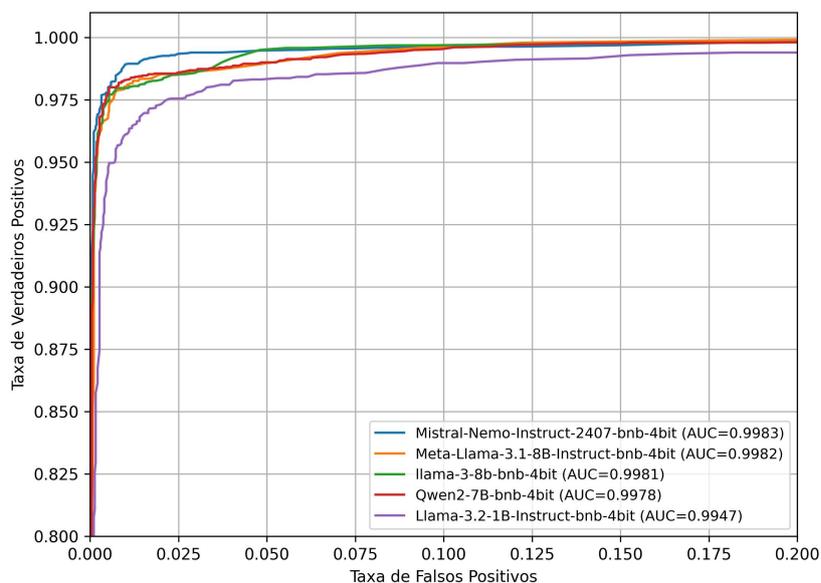
A Figura 15a apresenta as curvas ROC para todos os modelos. Observa-se que os modelos Mistral-Nemo, Llama-3.1-8B, Llama-3-8B, Qwen2-7B e Llama-3.2-1B apresentam curvas ROC que se aproximam do ponto (0, 1), indicando uma alta taxa de verdadeiros positivos e baixa taxa de falsos positivos. Em contraste, modelos como Phi-3.5-mini, Mistral-7B e Phi-3-medium-4k apresentaram desempenho significativamente inferior, com curvas mais próximas da linha de aleatoriedade e valores de AUC abaixo de 0,6.

Para uma melhor análise, a Figura 15b apresenta um recorte ampliado da curva ROC para os cinco modelos que obtiverem melhor desempenho na curva. Llama-3.1-8B

(AUC = 0.9982), Llama-3-8B (AUC = 0.9981) e Qwen2-7B (AUC = 0.9978) apresentaram um desempenho muito similar, com curvas ROC muito próximas do canto superior esquerdo do gráfico, indicando uma altíssima taxa de verdadeiro positivo (recall) mesmo em condições de baixa taxa de falso positivo.



(a) Curvas ROC de todos os modelos avaliados.



(b) Curvas ROC dos TOP 5 modelos (zoom na região de alto desempenho).

Figura 15 – Análise comparativa das curvas ROC entre todos os modelos (a) e os cinco com melhor desempenho (b).

O Mistral-Nemo (AUC = 0.9983) apresentou o melhor desempenho. Já o modelo

Llama-3.2-1B (AUC = 0.9947), embora tenha apresentado um bom desempenho, mostra uma pequena queda na região inicial da curva ROC, sugerindo que ele pode ter uma leve tendência a cometer mais falsos positivos em certos limiares de decisão.

Os modelos Mistral-Nemo, Qwen2-7B e Llama-3-8B destacam-se como os mais robustos na avaliação comparativa. O Mistral-Nemo apresenta a maior acurácia, precisão e F1-Score, aliadas a um AUC elevado, refletindo equilíbrio entre sensibilidade e especificidade. O Qwen2-7B sobressai com o maior recall e AUC próximo ao máximo, indicando eficácia em minimizar falsos negativos. O Llama-3-8B registra o terceiro AUC mais alto, demonstrando superior capacidade de discriminação entre classes, além de precisão e recall competitivos.

Modelos como Llama-3.1-8B e Llama-3.2-1B, embora apresentem AUC relevante (0,9982 e 0,9947, respectivamente), exibem acurácia e F1-Score inferiores (0,9853 e 0,9772), o que pode comprometer sua generalização. Em contraste, Gemma-2-9B, Mistral-7B e as variantes Phi-3 revelam falhas críticas: o primeiro tem F1-Score (0,6749) e precisão (0,5109) drasticamente reduzidos, enquanto os demais registram AUC próximo a 0,5 (0,3841 a 0,5298) e métricas igual a zero em precisão, recall e F1-Score, indicando desempenho equivalente a um modelo aleatório.

Na tarefa de classificação de similaridade entre marcas nominativas, é fundamental não apenas identificar corretamente os casos similares, mas também evitar falsas associações. Considerando a necessidade de equilíbrio entre a capacidade de identificar corretamente todas as marcas similares (recall) e a precisão em evitar classificações equivocadas (precisão), a métrica mais adequada para avaliar o desempenho dos modelos é o F1-Score, que representa a média harmônica entre precisão e recall.

Embora a ocorrência de falsos negativos seja particularmente crítica, pois pode permitir o registro de sinais potencialmente colidentes, a presença de falsos positivos também deve ser minimizada para evitar rejeições indevidas e onerar o processo de registro. Assim, a utilização do F1-Score como principal métrica de avaliação alinha-se à necessidade de assegurar a qualidade e a confiabilidade da triagem automatizada de marcas, equilibrando a severidade na identificação de riscos com a preservação da eficiência e da justiça no processo decisório.

Portanto, selecionamos o Mistral-Nemo, Qwen2-7B e Llama-3-8B, que tiveram melhor F1-score respectivamente, para avançar à etapa subsequente do método, dedicada à geração de relatórios explicativos sobre a similaridade entre marcas. Esses modelos foram escolhidos devido à sua consistência nas métricas e robusta capacidade discriminativa, indicando forte potencial para produzir relatórios precisos e alinhados às diretrizes do INPI.

6.2 Modelos Explicadores

Nesta seção, apresentamos a avaliação dos especialistas em relação ao desempenho dos modelos Llama3-8b, Qwen2 e Mistral-Nemo ao gerar relatórios considerando a explicação dos motivos ou não de colisão entre marcas similares e diferentes. Para avaliar o desempenho dos modelos, foram utilizadas as seguintes medidas estatísticas descritivas: média, mediana, moda e desvio padrão. Também foi realizada uma análise gráfica por meio de um boxplot. Esse gráfico permite visualizar de forma clara a dispersão dos dados, identificar outliers e comparar a distribuição dos valores entre os grupos.

Para fazer essa análise, foram convidados quatro especialistas em PI, cujos perfis seguem descritos a seguir.

- Especialista 1: Mestre e Doutor em Ciência da Propriedade Intelectual. Pesquisador na área de PI. Empreendedor na área de desenvolvimento tecnológico em PI.
- Especialista 2: Graduado em Direito e Análise e Desenvolvimento de Sistemas. Pós-graduado em Direito Público e Direito Empresarial. Especialista em análise de marcas, atua há 5 anos no INPI na área de avaliação de marcas, tendo exercido a função de presidente deste setor durante 3 desses anos.
- Especialista 3: Doutor e Mestre em Ciência da Propriedade Intelectual. Bacharel em Administração de Empresas e História. Professor e pesquisador em PI. Atua como consultor em PI, tendo contribuído no depósito de 75 pedidos de patentes, 46 Programas de Computador, 450 registros de marcas e 15 desenhos industriais. Certificado pelo INPI, Fórum Nacional de Gestores de Inovação e Transferência de Tecnologia (FORTEC), SEBRAE e a WIPO.
- Especialista 4: Empresário e consultor na área de PI. Membro do Conselho Consultivo do FORTEC. Membro da Associação Brasileira da Propriedade Intelectual (ABPI) e da Comissão de PI da OAB-SP. Graduado em Farmácia e Bioquímica. Pós-graduado em Administração de Empresas. Também é inventor, com patentes concedidas em vários países e criações premiadas e reconhecidas no Brasil e no exterior.

Modelos com menor desvio padrão indicam maior consistência nas avaliações. O modelo Mistral-Nemo apresentou o menor desvio padrão (0,93), sugerindo maior concordância entre os especialistas quanto à qualidade dos relatórios gerados. Por outro lado, o modelo *Llama3-8b* apresentou o maior desvio (1,13), evidenciando maior variabilidade nas avaliações recebidas. A Tabela 7 apresenta os resultados.

Observa-se que o modelo Mistral-Nemo obteve a maior média geral (4,25), seguido por Qwen2 (4,12) e Llama3-8b (4,05). Embora as diferenças entre as médias sejam pequenas.

Tabela 7 – Estatísticas das Avaliações dos Modelos.

Modelo	Média	Mediana	Moda	Desvio Padrão
Llama3-8b	4,05	4	5	1,13
Mistral-Nemo	4,25	4	5	0,93
Qwen2	4,12	4	5	1,07

Além disso, os valores de desvio padrão revelam diferenças relevantes quanto à consistência das avaliações.

Esses resultados são reforçados pela Figura 16, que apresenta as notas por modelo em formato de boxplot. Os resultados demonstram o Mistral-Nemo destacando-se como o mais eficaz, seguido pelo Qwen2 e, por fim, o Llama3-8b. O Mistral-Nemo apresenta a média geral de 4,25, a mais alta entre os modelos. Todos os modelos apresentam a mediana igual a 4.

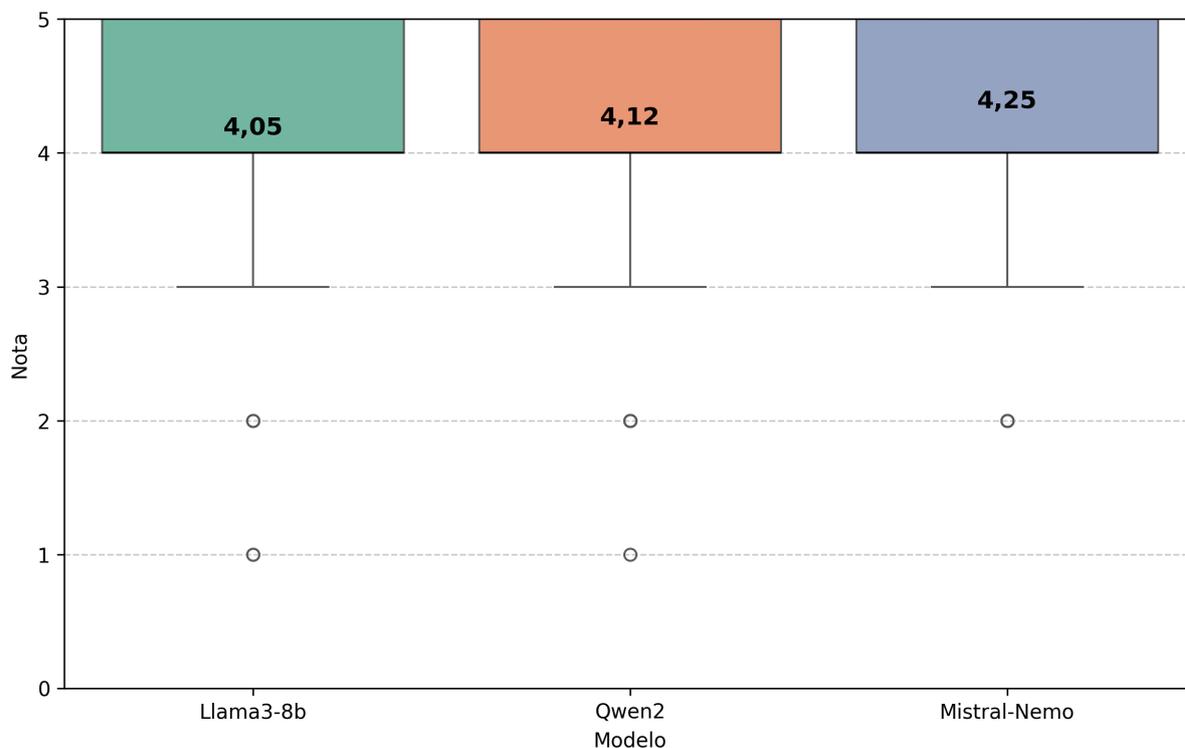


Figura 16 – Boxplot das Notas com Média Geral por Modelo.

Observa-se que o quartil superior (Q_3) não está visivelmente destacado em nenhum dos modelos analisados. Isso ocorre porque, nos três casos, o valor do terceiro quartil coincide com o valor máximo observado na distribuição das notas, isto é, $Q_3 = \text{máximo} = 5$. Em situações como essa, a linha superior da caixa do boxplot se sobrepõe ao topo do “bigode”, conferindo à visualização uma aparência de caixa truncada. Esse comportamento é estatisticamente esperado quando os dados apresentam baixa variabilidade nos valores superiores, com forte concentração de notas máximas.

Além disso, os círculos representados fora da caixa correspondem aos valores atípicos, que referem-se às notas baixas atribuídas aos modelos. Llama3-8b recebeu duas notas 1, três notas 2 e três notas 3; Qwen2 recebeu uma nota 1, quatro notas 2 e duas notas 3; e Mistral recebeu quatro notas 2 e uma nota 3. Esses pontos indicam observações que se desviam significativamente do padrão central da distribuição e são relevantes para identificar inconsistências, exceções ou erros cometidos pelos modelos.

A Figura 17 revela que os três modelos apresentam apenas pequenas variações em seu desempenho, tanto na avaliação de marcas consideradas semelhantes quanto na avaliação de marcas consideradas diferentes. Para relatórios similares, Llama3-8b, com média 4,15, e Mistral-Nemo, com média 4,20, enquanto o Qwen2 obteve uma média ligeiramente inferior (4,05). Nos relatórios para analisar marcas diferentes, o Mistral-Nemo mostrou-se superior (4,30), seguido pelo Qwen2 (4,20), enquanto o Llama3-8b teve desempenho menor (3,95).

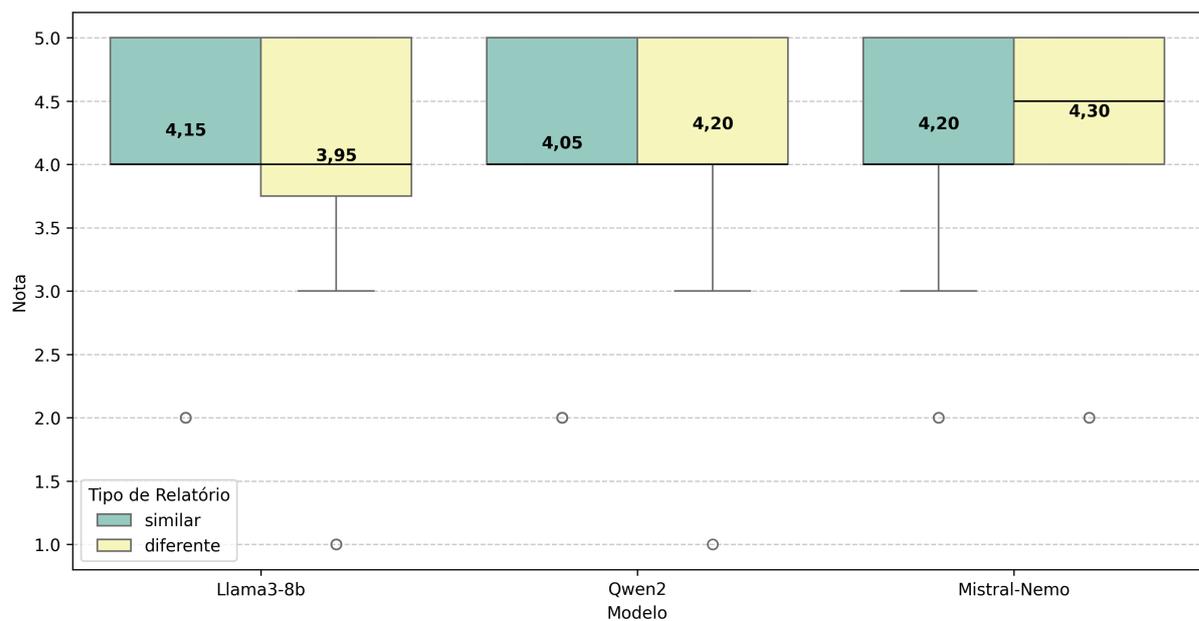


Figura 17 – Boxplot das Notas por Modelo e Tipo de Relatório.

O modelo Llama3-8b e o Qwen2 obtiveram mediana igual a 4 tanto nas avaliações de marcas semelhantes quanto nas de marcas diferentes. Já o modelo Mistral-Nemo apresentou mediana 4 nas avaliações de marcas semelhantes, mas alcançou uma mediana ligeiramente superior, de 4,5, nas avaliações de marcas diferentes.

6.2.1 Comentários dos Especialistas

Nesta seção, são apresentados os comentários deixados pelos especialistas em PI durante o processo de avaliação dos relatórios gerados pelos modelos explicadores. Observou-se que as anotações ocorreram, em sua maioria, nos casos em que os relatórios receberam notas mais baixas, geralmente acompanhadas de críticas ou sugestões de melhoria. No

entanto, também foram registrados alguns comentários em relatórios avaliados com nota máxima, nos quais os especialistas destacaram aspectos positivos, como a estrutura, clareza ou coerência da argumentação apresentada.

Um dos aspectos mais frequentemente observados pelos especialistas refere-se à limitação dos modelos em considerar apenas a classe NICE, sem se aprofundar na análise dos serviços ou produtos específicos associados à marca. O comentário apresentado na Figura 18a destaca a importância de avaliar a lista completa de produtos para uma identificação adequada de colidência. Essa preocupação é reiterada na Figura 18b, bem como nas observações das Figuras 18c, 18d e 18e, que reforçam a necessidade de uma contextualização mais precisa, enfatizando que a simples referência à classe NICE é insuficiente para uma avaliação robusta da afinidade mercadológica.

Outra categoria recorrente nos comentários diz respeito à dificuldade dos modelos em identificar corretamente os elementos verdadeiramente distintivos das marcas. O comentário exibido na Figura 20b aponta que o modelo analisado focou em termos genéricos, desconsiderando o termo que, na visão do especialista, constituía o núcleo da marca. A mesma limitação é evidenciada na Figura 19b, onde o especialista critica o uso de termos descritivos e comuns como base para a avaliação, ressaltando que esses não deveriam compor o núcleo distintivo da marca.

A capacidade dos modelos em organizar logicamente os argumentos também foi abordada pelos especialistas. No comentário da Figura 20a, destaca-se positivamente a estrutura do relatório produzido pelo modelo Mistral-Nemo, cuja sequência argumentativa foi considerada lógica, clara e coesa.

Por fim, alguns comentários registraram a efetividade dos modelos em identificar corretamente a similaridade entre marcas, demonstrando adequação à tarefa. Esse reconhecimento é ilustrado na Figura 21, onde o especialista valida a conclusão do modelo quanto à similaridade, justificando a nota máxima atribuída em razão da superioridade do relatório em comparação aos demais modelos avaliados.

6.3 Discussão

Com base nos resultados apresentados, procede-se à discussão, com o intuito de compreender as implicações dos achados e identificar as principais tendências e divergências. A análise será conduzida em duas etapas: na primeira, examinamos os resultados dos modelos classificadores; na segunda, analisamos os resultados dos modelos explicadores, com base nas avaliações realizadas por especialistas, a fim de verificar de que maneira os LLMs se adaptaram à tarefa de análise de marcas.

Comentário do especialista

Essa conclusão reforça a necessidade (eu diria até obrigação) de avaliar as marcas pela sua lista de produtos (no caso da classe 11) para se fazer uma avaliação adequada da colidência: “Como não há um direcionamento mercadológico evidente que sugira que ambas operam exatamente no mesmo nicho, a possibilidade de confusão entre os consumidores é reduzida”

(a) Comentário feito para relatório do Llama3-8b.

Comentário do especialista

Assinalar apenas a classe sem assinalar os serviços associados diretamente à marca requerida não é suficiente para uma boa avaliação.

(b) Comentário feito para relatório do Qwen2

Comentário do especialista

Eu considero novamente que seria necessária uma descrição e análise detalhadas das marcas de acordo com os serviços associados, não a classe generalizada.

(c) Comentário feito para relatório do Mistral-Nemo.

Comentário do especialista

A avaliação feita pela IA considerou os elementos nominativos das marcas para avaliar os produtos associados e não se aprofundou na afinidade mercadológica entre as marcas, o que é essencial.

(d) Comentário feito para relatório do Mistral-Nemo.

Comentário do especialista

A classe 35 é repleta de diferentes segmentos de atividades, por exemplo: supermercado, farmácia, posto de combustível, comércio de produtos óticos, serviços de Marketing, gestão de negócios, comercialização de roupas, sapatarias, etc. Assinalar apenas a classe sem assinalar os serviços associados diretamente à marca requerida não é suficiente para uma boa avaliação.

(e) Comentário feito para relatório do Llama3-8b.

Figura 18 – Comentários dos especialistas que ressaltam a importância da contextualização dos serviços associados às classes NICE para uma análise mais precisa da similaridade entre marcas.

6.3.1 Modelos Classificadores para Identificação de Similaridade

6.3.1.1 Análise do Desempenho dos Modelos

A análise dos padrões de erro revela que os modelos de melhor desempenho apresentam uma distribuição equilibrada entre falsos positivos e falsos negativos, o que indica uma capacidade consistente de identificar corretamente ambas as classes (“Similar” e “Diferente”). Essa simetria sugere que os modelos conseguem captar de forma eficaz

Comentário do especialista

Embora a conclusão aponte de forma correta a similaridade entre as marcas, eu acredito fielmente que o elemento central da marca é a palavra TOP, e, portanto, deveria ser esse utilizado para analisar a distintividade entre as marcas em tela. A palavra NET é um termo comumente utilizado no setor de atividade e já é reconhecido como um termo de uso comum para assinalar serviços de internet.

(a) Comentário feito para relatório do Qwen2

Comentário do especialista

Termos meramente descritivos do seguimento de atividade, como SHOES e MODA PRAIA, não precisariam, na minha opinião nem ser considerados na avaliação de marcas nominativas, porque nunca serão elementos de uso exclusivo, e portanto não fazem parte do núcleo das marcas em tela.

(b) Comentário feito para relatório Llama3-8b

Figura 19 – Comentários apontando limitações dos modelos na identificação dos elementos realmente distintivos das marcas.

Comentário do especialista

Gostei da estrutura utilizada para argumentar, o relatório seguiu uma sequência lógica da análise.

(a) Comentário feito para relatório do Mistral-Nemo

Comentário do especialista

Muito boa a clareza na organização das ideias, o relatório apresentou uma progressão coerente entre os pontos analisados, com argumentos bastante convincentes.

(b) Comentário feito para relatório do Qwen2.

Figura 20 – Comentários que destacam positivamente a estrutura lógica e coerência dos relatórios gerados pelos modelos.

Comentário do especialista

O relatório apontou de forma correta a similaridade entre as marcas. As palavras Net e FIBRA é um termo usualmente utilizado para se referir a serviços de internet. Justifico a nota 5 por ter apresentado um relatório melhor que as outras IA.

Figura 21 – Comentário feito para relatório do Mistral-Nemo.

as nuances entre marcas que colidem e aquelas que não apresentam risco de confusão, refletindo um bom grau de generalização.

Por outro lado, modelos como Mistral-7B, Phi-3-Medium e Phi-3.5-Mini que

apresentaram F1-score, precisão e recall igual a 0,0 indicam uma falha na adaptação do modelo ao contexto da tarefa. O modelo Gemma-2-9B que, apesar de apresentar alta taxa de recall, resultou em uma acurácia de apenas 0,515. Logo, o modelo conseguiu identificar corretamente a maioria dos casos da classe positiva, mas teve dificuldades em diferenciar as marcas realmente distintas, levando a um alto número de falsos positivos.

O baixo desempenho dos modelos Phi-3.5-Mini, Phi-3-Medium e Mistral-7B pode ser atribuído principalmente à sua arquitetura *decoder-only*, que, por ser otimizada para geração autoregressiva de texto, não possui uma representação explícita de toda a sequência adequada para tarefas de classificação. Essa limitação estrutural foi agravada pela modificação da cabeça de saída, restringida apenas aos tokens “Yes” e “No”, sem a adição de uma camada de classificação dedicada, o que dificultou a adaptação do modelo para decisões binárias precisas. Além disso, a máscara causal inerente a esses modelos impede a integração bidirecional do contexto, limitando a compreensão global necessária para distinguir corretamente entre classes. Em conjunto, a expectativa de geração sequencial e a ausência de uma arquitetura voltada para classificação direta comprometeram a capacidade desses modelos de aprender a tarefa proposta, mesmo com ajustes no esquema de treinamento.

O Llama-3.1, Llama-3-8B e Qwen2-7B que obteve o melhor resultado de AUC, indica que esses modelos têm uma capacidade de separar corretamente as classes “Similar” e “Diferente”, sendo confiáveis na classificação binária de marcas. Por outro lado, os modelos Mistral-7B e Phi-3-Medium, que apresentaram AUC inferior a 0,4, sugerem que suas previsões são inconsistentes e não diferenciam adequadamente as classes, tornando-os ineficazes para a tarefa de classificação de similaridade entre marcas. Esses resultados reforçam as observações feitas anteriormente, onde esses modelos falharam ao atribuir corretamente as classes, possivelmente devido a uma incapacidade de generalização para o contexto da tarefa.

6.3.1.2 Análise de Erros

Para proporcionar uma compreensão abrangente e representativa, essa discussão será orientada por uma análise dos erros cometidos. Em outras palavras, serão examinados os exemplos nos quais os modelos classificaram de forma divergente do rótulo, com o objetivo de identificar e compreender o raciocínio adotado durante a execução da tarefa proposta.

Os modelos Mistral-Nemo, Qwen2-7B e Llama-3-8b cometeram 96, 98 e 102 erros de classificação, respectivamente. A Tabela 8 apresenta algumas das instâncias em que ocorreram essas classificações incorretas.

Tabela 8 – Erros cometidos pelo modelo: comparação entre marcas, com rótulo e predição.

Marca A	Marca B	Rótulo	Predito
LLAMA-3-8b			
REALEZA	REISE	0 (<i>Similar</i>)	1 (<i>Diferente</i>)
LOOK	LUCKE	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
ZARA	zazcart	0 (<i>Similar</i>)	1 (<i>Diferente</i>)
SASCAR	zazcart	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
Jingo	GYMGO	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
ASUS . COLLECTION	AZOS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
ASUS ZenPower	AZOS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
cloudpay	NUVEMPAY	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
KROKE	CROC COXINHA	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
DEFYNA	Clínica De Fina	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
OSI	OZZY CONSTRUTORA	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
Reppi Hotéis	HAPPY HOTÉIS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
QWEN2-7b			
O Z	OS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
REALEZA	REISE	0 (<i>Similar</i>)	1 (<i>Diferente</i>)
ZARA	zazcart	0 (<i>Similar</i>)	1 (<i>Diferente</i>)
SASCAR	zazcart	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
Jingo	GYMGO	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
ASUS . COLLECTION	AZOS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
ASUS ZenPower	AZOS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
cloudpay	NUVEMPAY	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
elo	ELLO INVESTIMENTOS	0 (<i>Similar</i>)	1 (<i>Diferente</i>)
XIS DA DEDA	X ROOTS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
DEFYNA	Clínica De Fina	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
VYS FACTOR	VIZ BIOSCIENCE	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
Mistral-Nemo			
SYN	sim	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
BABY GAP	GUP	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
CIMM	SIM	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
Zama	SAMAS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
REALEZA	REISE	0 (<i>Similar</i>)	1 (<i>Diferente</i>)
ASUS ZenPower	AZOS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
cloudpay	NUVEMPAY	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
Jingo	GYMGO	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
ASUS TAICHI	AZOS	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
DEFYNA	Clínica De Fina	1 (<i>Diferente</i>)	0 (<i>Similar</i>)
OSI	OZZY CONSTRUTORA	1 (<i>Diferente</i>)	0 (<i>Similar</i>)

Continua na próxima página

Tabela 8 – Continuação

Marca A	Marca B	Rótulo	Predito
Reppi Hotéis	HAPPY HOTÉIS	1 (Diferente)	0 (Similar)
CAFÉ QUATRO AMORES	WITH LOVE PIACERE CAFFÈ	0 (Similar)	1 (Diferente)

Um exemplo ocorre ao verificarmos a análise feita entre as marcas *cloudpay* e *NUVEMPAY*, onde os três modelos divergiram ao classificar as marcas como “Similares” (rótulo 0), enquanto o rótulo verdadeiro é “Diferente” (rótulo 1). Essa predição, no entanto, revela uma compreensão válida por parte dos modelos quanto à similaridade ideológica entre os sinais. Ambas as marcas compartilham a estrutura “conceito digital + PAY”, sendo que “cloud” e “nuvem” são traduções diretas uma da outra em inglês e português, respectivamente, logo, ambas as marcas transmitem a mesma ideia. Percebemos assim o raciocínio gerado pelos modelos ao classificar essas marcas como semelhantes. Embora o critério de fonética aponte diferenças evidentes, como a ausência de sons comuns e variações no ritmo silábico, e a distintividade gráfica seja mantida por conta das raízes lexicais distintas, o modelo pode ter atribuído maior peso ao aspecto ideológico compartilhado. Ainda assim, a justificativa da predição revela um raciocínio desenvolvido pelos LLMs coerente com os critérios reais utilizados pelos especialistas em PI ao avaliar similaridade.

Na comparação entre as marcas *LOOK* e *LUCKE*, o modelo Llama-3-8b previu como “Similar” (rótulo 0), enquanto o rótulo verdadeiro é “Diferente” (rótulo 1). Novamente, a decisão do modelo é compreensível diante da elevada semelhança fonética e ortográfica dos sinais. Ambas compartilham o início “L” e o núcleo fonológico “U...K”, que pode soar de forma semelhante dependendo da pronúncia. O modelo, treinado exclusivamente em exemplos rotulados, parece ter identificado um padrão recorrente de similaridade com base nessas estruturas fonológicas. Contudo, os especialistas do INPI consideraram as marcas suficientemente distintas, possivelmente em função da terminação “E” e de outros fatores, como mercadológicos, não fornecidos no prompt.

Na comparação entre as marcas *Reppi Hotéis* e *HAPPY HOTÉIS*, o Llama-3-8b e Mistral-Nemo classificou-as como **Similares** (rótulo 0), enquanto o rótulo era **Diferente** (rótulo 1). Esta predição incorreta pode ser explicada principalmente pela elevada similaridade fonética entre os termos *Reppi* e *Happy*, que compartilham uma estrutura silábica muito próxima e sons consonantais e vocálicos similares. Tal proximidade fonética, mesmo sem uma correspondência ortográfica direta, é um dos critérios utilizados na análise de colidência entre marcas, e o modelo parece ter aprendido esse padrão com sucesso ao longo do treinamento.

Além disso, ambas as marcas contêm o termo “Hotéis”, que pode ter reforçado a percepção de afinidade mercadológica, sugerindo que atuam em segmentos semelhantes do setor de hospedagem. A junção desses fatores fonético-ideológicos pode ter induzido o

modelo a considerar o par como colidente. Portanto, reflete uma decisão compreensível por parte do modelo, que reconheceu padrões legítimos, mas não teve acesso ao contexto mercadológico mais amplo considerado na rotulagem oficial.

Logo, observa-se que esse mesmo padrão se repetiu em diversos casos de erro. Ao analisar as classificações, nota-se que a maioria decorreu do fato de os modelos considerarem aspectos fonéticos e ideológicos ao inferirem similaridade entre as marcas. Exemplos como *SASCAR* e *zazcart*, *Jingo* e *GYMGO*, *DEFYNA* e *Clínica De Fina*, bem como *J.J. NET* e *JOTA JOTA PODCAST*, demonstram forte sobreposição fonética — o que justifica a decisão do modelo, mesmo que ela tenha divergido do rótulo atribuído por especialistas.

Portanto, a partir dessa análise, observa-se que as classificações divergentes em relação aos rótulos atribuídos por especialistas não indicam, necessariamente, falhas por parte dos modelos LLMs. Ao contrário, os resultados sugerem que os modelos basearam suas decisões principalmente em critérios fonéticos e ideológicos — um padrão recorrente verificado em diversos exemplos da Tabela 8. Isso demonstra que, mesmo quando suas respostas diferem daquelas dos especialistas, os modelos não erraram de maneira arbitrária, mas sim seguiram uma lógica coerente com base nos dados e padrões fonológicos e ideológicos que internalizaram durante o treinamento.

Essa divergência também revela uma diferença fundamental no processo de tomada de decisão: os especialistas do INPI têm acesso a um conjunto muito mais amplo de informações contextuais (como classe Nice, setor de atuação, elementos gráficos e histórico de uso), o que naturalmente influencia sua avaliação de similaridade ou distinção entre marcas. Já os modelos foram treinados exclusivamente com nomes de marcas e os respectivos rótulos binários (“Similar” ou “Diferente”), sem qualquer acesso a contexto adicional.

Apesar dessa limitação, os modelos demonstraram uma notável capacidade de raciocínio autônomo. Mesmo sem instruções explícitas ou contextualização detalhada, os LLMs conseguiram identificar padrões fonéticos, morfológicos e ideológicos relevantes — como a presença de radicais compartilhados, estruturas sonoras semelhantes e significados afins — que, de fato, contribuem para a percepção de similaridade entre marcas. Essa habilidade emergente evidencia o potencial dos LLMs em aprender e aplicar critérios complexos apenas a partir de exemplos rotulados, capturando nuances sutis que muitas vezes guiam decisões humanas em tarefas de propriedade intelectual. Tal desempenho reforça a viabilidade do uso de LLMs como ferramentas auxiliares no processo de avaliação de colidência marcária, sobretudo em fases preliminares de triagem e análise automática.

6.3.2 Modelos Explicadores para Geração de Relatórios

A dispersão das notas sugere maior adaptação do Mistral-Nemo para realizar análises, com variação de 0,10 entre os tipos de relatórios, indicando estabilidade na

qualidade das respostas, independentemente do tipo de análise solicitada. Ao passo que o Llama3-8b apresentou maior frequência de avaliações baixas, com 8 notas iguais ou inferiores a 3, sinalizando desempenho ruim em algumas ocasiões, apesar de sua alta média geral, demonstrando maior especialização em tarefas de avaliar a similaridade, mas perdendo significativamente qualidade em relatórios de marcas diferentes, com variação de 0,20 entre os tipos de relatórios, sugerindo melhor adaptação do modelo ao gerar uma explicação da não colidência entre marcas diferentes. O Qwen2 ficou numa posição intermediária, mantendo médias altas nos dois cenários.

Observamos que nenhum dos modelos recebeu a nota 0 nas avaliações; a pontuação mínima observada foi 1, atribuída ao Llama e Qwen ao avaliar marcas diferentes.

Em termos de padrões de desempenho e consistência, os resultados indicam diferenças importantes. O Mistral-Nemo mostrou um desempenho muito consistente e robusto, já que evitou por completo as notas mínimas e concentrou suas avaliações nas categorias mais altas – um sinal de que suas respostas foram quase sempre de boa qualidade ou excelentes. O Qwen2 também demonstrou desempenho consistente majoritariamente positivo. O Llama3-8b, embora tenha obtido um volume considerável de notas máximas, revelou maior variabilidade nos resultados: além de várias avaliações excelentes.

A análise dos dados revela que o Mistral-Nemo não apenas obteve as médias mais altas, mas também a menor dispersão (caixas compactas e outliers ausentes), caracterizando-o como o modelo mais confiável para gerar relatórios técnicos em contextos jurídicos de PI. Essa consistência pode ser atribuída à sua capacidade de internalizar eficientemente os exemplos do prompt, replicando a estrutura argumentativa e a precisão terminológica observada nos casos de referência. Em contraste, a maior variabilidade do Llama3-8b em relatórios de diferença sugere dificuldade em adaptar-se a nuances contextuais quando a análise demanda contraposição explícita de elementos fonéticos, ideológicos ou mercadológicos.

A superioridade do Mistral-Nemo não apenas em média, mas também em consistência (menor Intervalo Interquartil (IQR)), reforça sua adequação para aplicações em ambientes regulatórios de PI, onde a confiabilidade é tão crítica quanto a precisão. A arquitetura do modelo, possivelmente mais adaptada à internalização de exemplos complexos via few-shot prompting, permite replicar estruturas argumentativas com menor dependência do tipo de relatório solicitado.

Em contraste, a maior dispersão do Llama3-8b sugere que, embora capaz de gerar análises de alta qualidade pontualmente, sua aplicação em cenários reais exigiria validação humana mais frequente para mitigar riscos de inconsistência. O Qwen2, por sua vez, representa um equilíbrio entre custo computacional e desempenho, sendo viável para contextos onde tolera-se uma margem moderada de variabilidade.

Logo, o modelo Mistral-Nemo revelou-se o mais interessante entre os modelos avaliados devido ao seu desempenho consistentemente elevado e baixa variabilidade nas respostas obtidas. A distribuição das suas avaliações demonstrou claramente uma forte concentração nas notas altas (principalmente 4 e 5), sem apresentar qualquer ocorrência das notas mais baixas (0 ou 1), que são indicativas de falhas críticas. Em comparação, embora o modelo Qwen2 tenha alcançado boas avaliações, ele ainda registrou ocorrência pontuais de falhas, demonstrando, portanto, uma leve oscilação. Já o modelo Llama3-8b destacou-se negativamente devido à maior dispersão de suas avaliações, alternando entre desempenhos muito altos e baixos, tornando-o menos confiável em cenários de uso real.

Os três modelos avaliados obtiveram, em geral, avaliações positivas pelos especialistas, indicando qualidade elevada na maioria dos relatórios explicativos gerados. Todos apresentaram médias de notas superiores a 4 numa escala de 0 a 5, com moda e mediana igual a 4. Isso significa que a maioria das avaliações foi nas faixas mais altas, refletindo consenso dos especialistas de que os relatórios estavam majoritariamente corretos e bem fundamentados. Em particular, o modelo Mistral-Nemo destacou-se com a maior média geral (4,25) e o menor desvio padrão (0,93), caracterizando um desempenho consistentemente bom. Sua distribuição de notas em boxplot apresenta uma caixa interquartil compacta e com somente 1 outliers, sugerindo alta confiabilidade nas explicações fornecidas. Essa estabilidade indica que o Mistral-Nemo manteve um nível de qualidade uniforme mesmo em cenários mais complexos, possivelmente devido à capacidade de internalizar eficientemente a estrutura argumentativa e a precisão terminológica dos exemplos fornecidos no prompt. Os especialistas reconheceram nesses relatórios uma redação alinhada às práticas do INPI, com uso adequado de termos técnicos e cobertura dos principais critérios (fonético, ideológico, distintividade e mercado).

O modelo Qwen2-7B também apresentou desempenho sólido, com média geral elevada (4,12) e mediana/moda em 4. Destacou-se por gerar um grande número de explicações excelentes, obtendo 18 ocorrências da nota máxima 5. A maioria (cerca de 82%) de suas avaliações foi de nota 4 ou 5, o que indica que, em grande parte dos casos, o Qwen2 forneceu relatórios muito bem avaliados pelos especialistas. Os comentários positivos frequentes mencionaram que suas respostas eram claras e estruturadas, muitas vezes seguindo fielmente os aspectos exigidos (por exemplo, comparando sistematicamente as marcas nos quatro pilares do INPI: aspecto fonético, ideológico, distintividade e afinidade mercadológica). Esse modelo mostrou ser capaz de balancear esses critérios de forma satisfatória na maioria das análises, o que se alinha com seu bom desempenho quantitativo. Além disso, o Qwen2 manteve desempenho alto tanto em relatórios de similaridade quanto de diferença, com médias sempre acima de 4. Essa versatilidade sugere um entendimento genérico da tarefa, conseguindo explicar corretamente por que duas marcas colidem ou não colidem na maioria dos cenários apresentados.

O Llama3-8b, embora tenha obtido a menor média dos três (4,05) , também apresentou moda e mediana de 4, indicando que grande parte de suas avaliações foi máxima ou próxima do máximo. 80% de suas notas ficaram em 4 ou 5, revelando que o modelo é capaz de produzir relatórios de altíssimo nível em muitas ocasiões. Os especialistas ressaltaram pontos positivos nos relatórios bem-sucedidos do Llama3-8b, como análises detalhadas e argumentação robusta quando o modelo acertava na abordagem. Em especial, nos casos de marcas similares, o Llama3-8b atingiu médias de desempenho comparáveis às do Mistral-Nemo (4,15) , sugerindo que ele possui forte habilidade para identificar e explicar conflitos quando as marcas de fato se assemelham. Nesses relatórios, os avaliadores elogiaram a forma como o modelo destacava os elementos coincidentes entre os sinais (e.g., raízes comuns ou fonéticas semelhantes) e articulava por que tais semelhanças causariam confusão do ponto de vista do consumidor. Tais casos indicam que, quando opera em seu melhor cenário, o Llama3-8b consegue replicar a lógica humana esperada, possivelmente graças à aprendizagem dos exemplos fornecidos (*few-shot learning*) e à incorporação correta das diretrizes do INPI.

Apesar do bom desempenho médio, os especialistas identificaram deficiências importantes e padrões de inconsistência em alguns relatórios, variando de modelo para modelo. O Llama3-8b apresentou o maior desvio padrão (1,13), apresentando uma maior quantidade de notas abaixo ou igual a 3. A análise detalhada revelou um padrão: as explicações mal avaliadas do Llama3-8b ocorriam majoritariamente nos relatórios onde deveria explicar a diferença entre as marcas (não colidência), nos quais sua média de notas caiu para 3,95, sensivelmente abaixo dos concorrentes. Logo, isso sugere dificuldades do Llama3-8b em justificar por que duas marcas não entram em conflito, não conseguindo capturar características que distinguem as marcas.

Comentários críticos apontaram que, em alguns desses relatórios, o modelo não deixou claro o porquê da distinção entre as marcas: por vezes descrevia características das marcas, mas não vinculava essas diferenças aos critérios de forma convincente, ou então omitia algum dos aspectos relevantes (por exemplo, focava em diferença fonética mas negligenciava a afinidade de mercado). Em outras situações, a argumentação apresentava contradições ou falta de profundidade, indicando que o modelo possivelmente “se perdeu” ao tentar aplicar as diretrizes do INPI em casos onde as marcas não possuíam semelhanças óbvias. Como podemos observar no comentário da Figura 18a:

Tais falhas resultaram em avaliações severas (notas 1) pelos especialistas, que nas suas observações enfatizaram a ausência de lógica consistente ou erro de interpretação das diretrizes nesses casos. Em suma, o Llama3-8b mostrou-se menos confiável e irregular: alternou entre explicações excelentes e outras claramente inadequadas, o que seria problemático num cenário de uso real sem supervisão constante.

O Qwen2-7B, embora bem avaliado em geral, não esteve isento de críticas. Seu

desvio padrão de notas (1,07) indica algumas flutuações – ainda que menores que as do Llama3-8b – e houve uma ocorrência de nota 1 em suas avaliações. Os especialistas relataram que, pontualmente, alguns relatórios do Qwen2 careciam de detalhes ou precisão esperada. Em certos casos, o modelo generalizou em excesso a explicação, reaproveitando argumentos dos exemplos sem adaptá-los perfeitamente ao caso concreto. Por exemplo, um avaliador comentou que um relatório apresentava uma “estrutura correta, porém com análises superficiais”, quase como se o modelo tivesse “seguido a fórmula” sem realmente captar a nuance específica das marcas avaliadas. Como visto no comentário da Figura 20b

Contudo, apesar da crítica ao foco do relatório, o especialista reconhece que a conclusão foi correta, o que denota clareza e estrutura válida no parecer. Outro comentário indicou que, embora o Qwen2 mencionasse todos os aspectos, faltou evidenciar com clareza a distinção mercadológica em um caso de marcas diferentes – sinal de que o modelo talvez não tenha compreendido totalmente a importância de citar a diferença nos produtos/serviços associados (classes de Nice) para justificar a ausência de conflito. Como observado no comentário da Figura 18b

Ainda assim, vale ressaltar que tais deficiências do Qwen2 foram relativamente raras. A maior parte de suas explicações foi bem recebida, e as críticas se concentraram em melhorias incrementais (por exemplo, sugestões de incluir exemplos adicionais ou reforçar determinado argumento) ao invés de erros graves. Isso se reflete na distribuição de notas: o Qwen2 teve predominância de avaliações 4 e 5, e apenas alguns casos nos níveis baixos, indicando oscilações menores e menos frequentes em comparação ao Llama3-8b.

O Mistral-Nemo apresentou pouquíssimas deficiências evidentes segundo os especialistas. Seu desempenho foi não só o melhor em média, mas também o mais consistente, sem nenhuma nota extremamente baixa registrada (nenhum 0 ou 1). Consequentemente, suas explicações raramente receberam críticas severas. No entanto, pequenas observações de melhoria também foram apontadas. Em um ou outro relatório, avaliadores notaram que o Mistral-Nemo poderia aprofundar um pouco mais a análise ideológica – por exemplo, explicitar com mais clareza por que os conceitos evocados por duas marcas similares causariam confusão no público, como visto no comentário da Figura 18c.

Em geral, as falhas do Mistral foram sutis: eventualmente repetição de argumentos já mencionados (talvez por zelo em reforçar o ponto), ou falta de síntese em alguns trechos, tornando o relatório mais longo do que o necessário. Tais aspectos não comprometeram gravemente a qualidade, mas indicam espaço para refinamentos pontuais. Os especialistas parecem ter tido alta confiança nesse modelo, dado que mesmo suas notas mínimas caíram na faixa intermediária (não recebendo nota 0 ou 1). Assim, o padrão negativo – se é que pode ser chamado assim – do Mistral-Nemo foi essencialmente “a ausência de falhas críticas”, com apenas ajustes marginais sugeridos.

Um ponto observado pelos especialistas foi a capacidade dos modelos em se ade-

quarem às diretrizes oficiais do INPI. De modo geral, os avaliadores esperavam que os relatórios seguissem de perto os critérios padronizados (fonético, ideológico, distintividade e afinidade mercadológica), aplicando-os de forma consistente e fundamentada a cada par de marcas. Quando os modelos cumpriram esse objetivo, os comentários tenderam a ser elogiosos ou neutros. Por exemplo, foi observado que vários relatórios do Mistral-Nemo e do Qwen2 “espelharam a estrutura esperada” isto é, examinaram primeiramente semelhanças/diferenças fonéticas (sons, entonação, sílabas coincidentes), depois avaliaram se os nomes evocavam ideias ou conceitos semelhantes (critério ideológico), identificaram elementos distintivos dominantes (como prefixos ou termos únicos) e, por fim, compararam as áreas de atuação ou classes Nice de cada marca, chegando a uma conclusão ao final, como observado pelo comentário da Figura 20a

Por outro lado, os comentários críticos lançam luz sobre casos em que os modelos aparentaram “não compreender plenamente alguma nuance das diretrizes”. Uma crítica recorrente foi a omissão ou tratamento inadequado de um dos quatro aspectos: em algumas explicações de diferença (marcas não colidentes) que receberam baixa nota, os especialistas notaram que o modelo “deixou de analisar o aspecto X” – por exemplo, não discutiu a afinidade mercadológica quando as classes de produtos eram claramente distintas, ou não mencionou o critério ideológico num caso onde as marcas possuíam temáticas totalmente diferentes. Este comportamento pode ser observado no comentário da Figura 18d. Isso sugere que o modelo, em alguns casos, não internalizou completamente a necessidade de examinar todos os ângulos antes de concluir pela ausência de conflito.

Em outros casos, a crítica foi sobre interpretações equivocadas: houve relatórios onde o modelo afirmou que duas marcas eram suficientemente diferentes por causa de um detalhe trivial, ou inversamente, alegou similaridade forte com base em um ponto fraco. Um avaliador, ao revisar uma explicação mal-sucedida do Llama3-8b, observou que “o relatório repete termos técnicos, mas não parece entender o caso concreto – falta justificar com propriedade”. Esse tipo de comentário indica que o modelo pode ter aprendido a “forma” das justificativas (a linguagem e estrutura), porém em certos casos falhou em aprender o “conteúdo substancial”, isto é, quando e como cada critério é determinante para o resultado. Como observado no comentário da Figura 19b. Os especialistas perceberam nesses relatórios uma compreensão superficial ou literal dos exemplos de prompt, sem verdadeiro entendimento contextual.

Os especialistas forneceram sugestões valiosas que revelam lacunas recorrentes nos modelos. Uma observação frequente foi a necessidade de contextualizar melhor as diferenças mercadológicas: por exemplo, ao lidar com marcas que operam em setores distintos (classes de Nice diferentes), o modelo deveria enfatizar que essa divergência reduz o risco de confusão. A ausência dessa consideração sugere que alguns relatórios não atribuíram peso suficiente ao fator mercadológico, possivelmente por limitações na

compreensão do contexto de uso das marcas — elemento central nas diretrizes do INPI. Essa observação fica evidente no comentário da Figura 18e

Outra recomendação foi evitar argumentos excessivamente genéricos, como afirmar que “os nomes são diferentes” sem justificar tecnicamente essa diferença. Os especialistas destacaram a importância de sempre vincular a distinção a critérios concretos, como fonética, significado, elementos distintivos ou setor de atuação. Quando isso não ocorreu, os avaliadores perceberam falta de profundidade analítica, evidenciando uma compreensão incompleta das boas práticas em exames de colidência. A correlação entre fragilidades nas explicações e erros de classificação revela aspectos importantes sobre a lógica interna dos modelos. Na etapa de classificação binária — realizada sem explicações — os modelos Mistral-Nemo, Qwen2-7B e Llama3-8b atingiram altos índices de acurácia e F1-score, com curvas ROC quase perfeitas. Ainda assim, todos cometeram erros pontuais na predição de similaridade entre marcas. A análise desses erros mostra um alinhamento estreito com os padrões qualitativos apontados pelos especialistas.

O modelo Mistral-Nemo, por sua vez, demonstrou maior coerência entre explicação e classificação. Foi o que cometeu menos erros absolutos na classificação, onde nos relatórios gerados, que mesmo quando errou, havia justificativas plausíveis. Também foi o modelo que mais se adequou as estruturas da análise e com argumentos mais convincentes, como observado no comentário da Figura 21, ao avaliar as marcas “NETFIBRA TELECOMUNICAÇÕES” e “Net FIBRA”. Isso reflete o raciocínio lógico que poderia ser compartilhado por um examinador humano, diferentemente de erros mais desconexos observados no Llama3-8b, cujas explicações receberam inclusive notas 1 por falta de aprofundamento.

Portanto, a ausência de explicações extremamente frágeis no Mistral-Nemo está em consonância com sua ausência de erros grosseiros na classificação. Ambos os aspectos sinalizam um modelo com entendimento mais completo da tarefa, capaz de cometer apenas equívocos pontuais em casos ambíguos. Essa convergência entre o desempenho explicativo e classificatório reforça a utilidade da explicação como indicador da lógica interna do modelo e oferece subsídios concretos para seu aprimoramento.

Em síntese, os pontos fortes gerais incluem a estruturação coerente dos relatórios — com introdução do caso, análise dos critérios e conclusão —, o uso apropriado de linguagem técnico-jurídica e a consideração explícita dos quatro aspectos de similaridade exigidos. Os modelos que mais se destacaram, Mistral-Nemo e Qwen2, apresentaram explicações consistentes e adequadas, demonstrando capacidade de generalização ao lidarem com casos de marcas colidentes e não colidentes com variações mínimas de desempenho.

Por outro lado, os padrões negativos observaram inconsistências relevantes no Llama3-8b, leves omissões ocasionais no Qwen2 e raras oportunidades de refinamento no Mistral-Nemo. Essas deficiências estiveram, em geral, associadas à dificuldade dos modelos em equilibrar os critérios exigidos pelo INPI ou à falta de clareza lógica na vinculação entre

fatos e conclusões. Cada modelo, assim, apresenta um perfil de risco distinto: Llama3-8b, com maior propensão a erros graves esporádicos; o Qwen2, com omissões pontuais; e o Mistral-Nemo, com desempenho estável e confiável neste conjunto de testes.

De forma geral, os comentários dos especialistas indicam que os modelos assimilaram a “linguagem” das diretrizes — isto é, apresentam argumentos dentro do formato esperado —, mas ainda não alcançaram a “profundidade decisória” de um examinador humano. Em alguns casos, os modelos construíram justificativas formais, porém não identificaram corretamente quais aspectos deveriam prevalecer em determinada situação, o que comprometeu a conclusão. Isso revela que as falhas não são meramente classificatórias, mas estão associadas a desequilíbrios internos na ponderação dos critérios utilizados.

Entre os aspectos analisados, destaca-se o fator mercadológico — especialmente relacionado ao setor de atuação das marcas — como elemento decisivo na avaliação de risco de confusão. As críticas apontaram, com frequência, a ausência de uma análise aprofundada sobre os serviços ou produtos efetivamente associados às marcas. Isso reforça a importância de incluir esse critério de maneira explícita no treinamento dos modelos explicadores, sobretudo por meio da incorporação de informações sobre o nicho de mercado e da interpretação adequada da Classe Nice. Tal abordagem pode contribuir para explicações mais completas e alinhadas às diretrizes reais adotadas nos exames técnicos.

A relação entre as métricas quantitativas e qualitativas mostrou-se coerente: os modelos que obtiveram melhores notas explicativas também apresentaram maior acurácia na classificação. Quando um modelo fornece justificativas bem estruturadas e abrange todos os critérios relevantes, tende a chegar à decisão correta. Por outro lado, explicações incompletas ou logicamente frágeis indicam, em geral, decisões equivocadas. Essa coerência interna reforça a hipótese de que a explicação pode funcionar como uma janela para a “linha de raciocínio” do LLM, permitindo identificar deficiências específicas — como o desequilíbrio entre critérios fonético e mercadológico no Llama3-8b ou a superficialidade ocasional no Qwen2. Assim, os dados discutidos não apenas comparam desempenhos, mas também elucidam relações causais entre as escolhas explicativas dos modelos e seus respectivos acertos ou erros, oferecendo direções claras para seu aprimoramento futuro.

7 Considerações Finais

O objetivo geral deste trabalho foi propor um método baseado em LLMs para a análise de similaridade entre marcas nominativas. Esse método contempla dois componentes: um modelo classificador, treinado de forma supervisionada para identificar possíveis colisões entre marcas conforme os critérios técnicos adotados pelo INPI, e um modelo explicador, configurado por meio de engenharia de prompt, capaz de gerar relatórios textuais justificando as decisões do classificador sob os aspectos fonético, ideológico, distintivo e mercadológico. Ao longo desta dissertação, esse objetivo foi efetivamente alcançado por meio do desenvolvimento da solução proposta: construímos um conjunto de dados com casos reais de conflitos entre marcas, realizamos o fine-tuning de LLMs de código aberto para a tarefa de classificação binária de similaridade, e exploramos a geração de explicações alinhadas aos critérios exigidos.

Os resultados obtidos nos experimentos confirmam a efetividade da abordagem. Os modelos classificadores apresentaram desempenho promissor, com acurácia superior a 0,98 e AUC próxima de 0,99, indicando alta capacidade de distinguir similaridade entre marcas. Paralelamente, os modelos explicadores demonstraram se adaptar com aptidão em produzir relatórios detalhados e corretos: em grande parte dos casos, as justificativas geradas cobriram explicitamente os quatro critérios do INPI conforme solicitado no prompt, utilizando linguagem técnico-jurídica adequada e uma estrutura lógica (introdução do caso, análise de cada critério, conclusão). Um destaque específico foi o modelo Mistral-Nemo, que emergiu como a solução de melhor desempenho. Esse modelo não apenas obteve os melhores índices de classificação (com leve vantagem no F1-score em relação a outros LLMs testados), como também produziu explicações de forma consistente e confiável alinhadas ao prompt utilizado, sem cometer erros grosseiros.

Observou-se que, mesmo quando o Mistral-Nemo incorreu em equívocos pontuais, estes ocorreram em situações desafiadoras e seguindo uma lógica compreensível (por exemplo, marcas com nomes quase idênticos no mesmo setor de mercado). Assim, os experimentos evidenciam que a solução proposta – integrando modelos classificadores e explicadores – atende plenamente ao objetivo inicial: eles são capazes de identificar conflitos entre marcas com precisão e, simultaneamente, fornecer uma fundamentação transparente para cada decisão.

Esses resultados não apenas validam a eficácia técnica da solução proposta, como também fornecem evidências empíricas robustas para a confirmação das hipóteses de pesquisa estabelecidas nesta dissertação. Logo, esses resultados comprovam a primeira hipótese de pesquisa levantada, ao demonstrarem que os LLMs alcançaram elevado desempenho na

tarefa de classificação reconhecendo padrões fonéticos, ideológicos e distintivos mesmo sem o fornecimento de contexto adicional. A análise dos erros reforça essa conclusão, uma vez que os equívocos ocorreram, em sua maioria, em casos ambíguos e passíveis de interpretação subjetiva, nos quais os modelos ainda assim apresentaram justificativas coerentes. Esses achados confirmam que os LLM, quando adaptados com dados especializados, são capazes de replicar — e em alguns casos até antecipar — o raciocínio utilizado por especialistas em PI, validando a hipótese proposta nesta dissertação.

Adicionalmente, os resultados também confirmam a segunda hipótese de pesquisa. Os LLMs avaliados, após fine-tuning com dados reais de marcas brasileiras, não apenas demonstraram alto desempenho na tarefa de classificação de similaridade, como também foram capazes de gerar relatórios explicativos considerados satisfatórios por especialistas em PI. Dessa forma, confirma-se que um LLM treinado no contexto brasileiro pode, sim, não apenas realizar classificações automatizadas, mas também justificar suas decisões de forma compreensível e tecnicamente fundamentada. Isso representa um avanço importante no sentido de promover maior transparência, consistência e eficiência no processo de análise de colidência entre marcas, reduzindo a subjetividade inerente às avaliações exclusivamente humanas.

Conforme postulado no início do trabalho, LLMs devidamente especializados conseguem sim identificar similaridade entre marcas nominativas de maneira precisa e fundamentada. Os resultados mostraram que um LLM, quando treinado com dados reais de colisões de marcas, aprende as nuances desses critérios e alcança desempenho próximo ao ideal na classificação de conflitos. Além disso, ficou demonstrado que tais modelos podem ir além da simples decisão binária: por meio de engenharia de prompt e ajustes finos, os LLMs foram capazes de gerar relatórios explicativos detalhados que dão suporte às suas decisões. Essas explicações fornecem maior transparência e justificativa, reproduzindo em boa medida o racional que um examinador humano empregaria para aprovar ou rejeitar um pedido de registro.

Desse modo, a pesquisa não apenas atendeu aos objetivos propostos, como também valida empiricamente a premissa de que ferramentas de IA, especificamente LLMs open-source de porte reduzido (com até 8 bilhões de parâmetros) e adaptados por meio de *fine-tuning* com dados especializados, podem auxiliar e até aprimorar o processo de análise de anterioridade de marcas. Em suma, conclui-se que a abordagem desenvolvida representa um avanço significativo: integra alto desempenho com interpretabilidade, reduz a subjetividade nas avaliações ao explicitar os critérios considerados e tem potencial para aumentar a eficiência no exame de marcas, corroborando todas as expectativas levantadas na hipótese inicial.

7.1 Contribuições Científicas, Tecnológicas e Sociais

Do ponto de vista científico, a principal contribuição reside na introdução de um método baseado em LLM para a tarefa de avaliação de similaridade entre marcas, algo inédito na literatura até o momento. Diferentemente dos estudos anteriores, que em geral restringiam-se a aspectos isolados e não proviam mecanismos explicativos, o presente trabalho demonstrou que é possível abranger, em um só modelo, múltiplos aspectos de análise (fonético, ideológico, distintivo e mercadológico) e ainda conseguir adaptar modelos através de engenharia de prompt para gerar explicações compreensíveis. A solução proposta preenche, portanto, uma lacuna no estado da arte em IA aplicada à PI, ao fornecer um método completo de decisão e justificação.

Ademais, os experimentos conduzidos reforçam achados importantes para a comunidade acadêmica: comprovou-se que LLMs de código aberto relativamente pequenos (≤ 8 bilhões de parâmetros), quando treinados em um domínio específico, podem atingir performance satisfatória aos especialistas em PI. Esse resultado incentiva pesquisas futuras a explorar modelos open-source em cenários jurídicos, promovendo assim o desenvolvimento de tecnologias acessíveis e auditáveis nesse campo.

Do ponto de vista tecnológico, destaca-se primeiramente a criação e disponibilização do dataset CMAD. Trata-se de um conjunto de dados inédito, compilado a partir de casos reais de marcas conflitantes registrados no Brasil, com curadoria e anotação realizadas em conformidade com as decisões do INPI. O CMAD, contendo mais de dezoito mil exemplos de conflitos entre marcas, configura-se como um recurso valioso para a pesquisa e o desenvolvimento de sistemas de IA no domínio de marcas. Sua contribuição tecnológica é dupla: primeiramente, ele viabilizou o treinamento e a avaliação dos modelos nesta dissertação sob condições realistas, garantindo que o aprendizado dos LLMs estivesse alinhado às situações encontradas na prática. Em segundo lugar, ao ser disponibilizado para a comunidade (inclusive com a publicação de um artigo dedicado descrevendo-o), o CMAD passa a servir de base para que outros pesquisadores repliquem experimentos, comparem algoritmos e criem novas soluções de apoio à decisão em PI.

O método implementado representa a principal contribuição tecnológica desta pesquisa, ao demonstrar, de forma prática como modelos de linguagem podem ser integrados ao fluxo de análise de marcas. Essa integração abrange desde a etapa de preparação e especialização dos modelos, por meio de técnicas de fine-tuning com dados reais e representativos do contexto brasileiro, até a aplicação de estratégias eficazes de engenharia de prompt para a geração de explicações técnicas alinhadas aos critérios do INPI. Tais estratégias oferecem um guia replicável e adaptável para pesquisadores e desenvolvedores que desejam incorporar LLM em sistemas de apoio à decisão, sejam eles voltados à triagem automática, à elaboração assistida de relatórios técnicos ou à análise preliminar de risco de colidência. Assim, além de contribuir diretamente com uma solução funcional, o método

proposto fornece um arcabouço que pode ser reaproveitado ou expandido em aplicações futuras no domínio da PI e em outras áreas jurídicas e administrativas que demandem análises complexas baseadas em texto.

Por exemplo, escritórios de patentes e marcas ou empresas de tecnologia jurídica podem se beneficiar das estratégias de engenharia de prompt aqui propostas, especialmente na geração automatizada de relatórios explicativos. Embora a replicação do modelo classificador demande infraestrutura adequada e dados rotulados, os resultados obtidos indicam caminhos promissores para o desenvolvimento de assistentes inteligentes no apoio à análise de colidência entre marcas.

Por fim, sob o prisma social, a modernização do processo de exame de marcas, por meio de ferramentas de IA, pode tornar o registro de marcas mais ágil, acessível e seguro. Pequenos empreendedores e startups, que muitas vezes não dispõem de recursos para conduzir buscas extensivas de anterioridade, poderão ser beneficiados indiretamente por tecnologias como a aqui desenvolvida – seja via futuros serviços online que utilizem modelos treinados para alertar sobre possíveis conflitos antes do depósito de um pedido, seja via um aprimoramento do próprio INPI na triagem de pedidos colidentes. Ao reduzir o tempo de análise e minimizar o risco de decisões inconsistentes, um sistema automatizado de apoio pode aumentar a confiança no processo de registro de marcas e estimular a proteção de ativos intangíveis. Isso pode contribuir para um ambiente de negócios mais estável, no qual empresas têm maior segurança jurídica ao investir em marcas, e consumidores são menos expostos a confusão entre produtos/serviços similares.

Ademais, a transparência proporcionada pelos relatórios explicativos gerados pelo modelo tem um valor social importante: as partes envolvidas em um eventual conflito poderiam compreender melhor as razões de uma decisão (por exemplo, por que seu pedido foi negado em virtude de colisão com marca anterior), o que favorece a clareza e possivelmente diminui a litigiosidade decorrente de inconformismo com decisões opacas.

7.2 Limitações

Apesar dos resultados promissores e das contribuições destacadas, é importante reconhecer algumas limitações deste estudo, sobretudo de natureza metodológica, que delimitam o escopo e sugerem cautela na generalização dos achados.

- A primeira diz respeito ao conjunto de dados utilizado. O número de exemplos do CMAD refletem as decisões e registros do INPI em um dado período de tempo. O que pode significar que o modelo aprendeu com base em exemplos que não são absolutamente homogêneos em critérios – por mais que as diretrizes do INPI balizem as decisões, sempre há margem para interpretação.

- Outra limitação reside na avaliação das explicações geradas pelo modelo. Diferentemente da classificação binária, cujo desempenho pôde ser medido objetivamente via métricas como acurácia, F1 e AUC, a avaliação de explicações textuais requer julgamentos qualitativos. Neste trabalho, optou-se por analisar as justificativas produzidas pelos LLMs à luz de critérios pré-definidos (cobertura dos quatro aspectos, correção técnica, coerência e clareza) e atribuir notas ou observações conforme o cumprimento desses requisitos. Ainda que tenhamos seguido um protocolo cuidadoso para essa avaliação (buscando identificar omissões, incoerências e acertos em cada relatório gerado), trata-se inevitavelmente de um processo subjetivo.

7.3 Trabalhos Futuros

A partir dos resultados obtidos neste trabalho, abre-se um caminho promissor para futuras pesquisas na área de aplicação de LLM à PI. Os achados aqui apresentados demonstram o potencial de modelos ajustados para auxiliar na análise de marcas, e indicam diversas possibilidades de aprofundamento, como a adaptação do método para marcas figurativas ou mistas, o uso de abordagens multimodais, e a expansão do conjunto de dados com novos casos de colidência.

Uma direção promissora está em explorar modelos de linguagem ainda mais recentes e robustos, com mais parâmetros. O campo de LLMs evolui rapidamente, e novos modelos open-source de maior porte ou melhor treinados (por exemplo, sucessores do Mistral-Nemo, versões aprimoradas do LLaMA, ou outros modelos emergentes) poderão ser incorporados à abordagem. Avaliar o método proposto com esses modelos de última geração pode elevar ainda mais a acurácia da classificação e a qualidade das explicações, além de verificar a generalização dos resultados obtidos.

Igualmente, seria interessante investigar técnicas de aperfeiçoamento do modelo explicador, como o uso de aprendizado por Reinforcement Learning from Human Feedback (RLHF) (do português, aprendizado por reforço com feedback humano) para alinhar as justificativas geradas aos padrões esperados por especialistas, tornando-as ainda mais fiéis às diretrizes oficiais. Outra linha de expansão envolve a adoção de abordagens multiagente ou com comitês de modelos. Em vez de depender de um único LLM para julgar e explicar, pode-se conceber um sistema composto por vários agentes inteligentes cooperando na tarefa. Por exemplo, diferentes modelos (ou instâncias do modelo) poderiam especializar-se em cada critério de similaridade: um agente focado na análise fonética, outro dedicado à comparação ideológica (conceitual), outro avaliando distintividade e mercado. Em seguida, um agente integrador consolidaria as avaliações parciais e produziria a decisão final e o relatório unificado.

Alternativamente, poderia-se empregar um agente verificador, que recebesse a

decisão e explicação propostas por outro modelo e as revisasse criticamente, apontando eventuais incoerências ou lacunas (um estilo de autoavaliação colaborativa). Abordagens multiagente como essas têm potencial para aumentar a confiabilidade do sistema, reduzindo erros e equilibrando melhor os diversos aspectos da decisão, pois simulam um processo de *peer review* automatizado entre modelos. Pesquisas futuras podem avaliar o custo-benefício dessa complexidade extra e comparar seu desempenho com o do modelo único atual.

Um terceiro eixo de trabalhos futuros diz respeito à integração do método em sistemas reais do INPI ou de escritórios de marcas. Tendo demonstrado viabilidade em ambiente controlado, o passo seguinte natural é levar a tecnologia para um contexto de uso prático. Isso envolveria desafios adicionais, como a necessidade de tratar fluxos contínuos de novos pedidos de marcas em tempo real, garantir tempos de resposta reduzidos e interfaces adequadas para que examinadores interajam com as sugestões do modelo. Uma possibilidade seria implementar um projeto-piloto em colaboração com o INPI, onde o classificador de similaridade e o gerador de relatórios atuassem como um sistema de apoio ao exame: a cada novo pedido depositado, o sistema poderia listar possíveis marcas conflituosas já registradas (algo similar ao que hoje é feito manualmente na busca de anterioridade) e fornecer um relatório preliminar indicando se o novo sinal provavelmente colidiria com alguma marca vigente, embasando sua conclusão nos quatro critérios. O examinador humano então utilizaria essa informação como subsídio, validando ou complementando a análise da máquina.

Estudos futuros também poderiam medir o impacto de tal ferramenta no tempo de processamento dos pedidos e na qualidade/consistência das decisões finais. Além disso, a integração real alimentaria um ciclo de retroalimentação: à medida que novas decisões de conflito são tomadas (com ou sem o auxílio da IA), esses casos poderiam ser incorporados ao conjunto de treinamento, permitindo um aprendizado contínuo do modelo e adaptação a eventuais mudanças de jurisprudência ou tendências de nomes no mercado (NGUYEN *et al.*, 2012). Por fim, enfatiza-se a importância de expandir e aprimorar o dataset para pesquisas subsequentes. O CMAD, em sua versão atual, poderá ser complementado com novos dados, aumentando tanto a quantidade quanto a diversidade dos exemplos. Uma primeira extensão seria incluir casos adicionais de anos mais recentes, acompanhando a evolução do registro de marcas no Brasil e capturando novos conflitos que surgirem.

Outra melhoria seria enriquecer o conjunto de exemplos negativos (marcas que coexistem sem conflito) de forma mais sistemática, talvez mapeando marcas registradas em uma mesma classe nice que não geraram oposição ou colisão, para fornecer contra-exemplos robustos ao classificador. Também seria valioso incorporar informações sobre o desfecho dos conflitos (por exemplo, decisões judiciais posteriores, acordos ou convívio pacífico) caso disponíveis, para permitir ao modelo distinguir graus de similaridade ou situações em que exceções ocorreram. Em um escopo mais amplo, futuras pesquisas podem almejar

a criação de um dataset multinacional de conflitos marcários, unindo dados de outros escritórios de PI (por exemplo, USPTO, CNIPO) ao CMAD.

Isso possibilitaria treinar modelos multilíngues ou sensíveis a diferentes legislações, ampliando o alcance da solução para contextos globais e testando sua adaptabilidade. Por último, mas não menos importante, a extensão do presente trabalho a marcas não puramente nominativas seria um caminho desafiador e recompensador. Logo, como trabalhos futuros pode-se adotar LLMs multimodais capaz de integrar dados de diversos tipos (imagem, texto, etc) para abranger todos os aspectos das marcas no momento de uma análise de similaridade, isso abriria a porta para um sistema unificado capaz de avaliar similaridade em marcas compostas (mistas figurativas/nominativas), cobrindo todo o espectro de propriedades distintivas que uma marca pode ter.

7.4 Escrita e Publicações de Artigos Científicos

Durante o período de pesquisa deste mestrado, foi possível elaborar os seguintes trabalhos científicos:

1. Estudo 1 (publicado): Reis, I.B.; Leite, R.Â.S.; Torres, M.M.; Neto, A.G.d.S.; Silva, F.J.d.S.e.; Teles, A.S. *Conflicting Marks Archive Dataset: A Dataset of Conflicting Marks from the Brazilian Intellectual Property Office*. Data 2024, 9, 33. <https://doi.org/10.3390/data9020033> (REIS *et al.*, 2024).
2. Estudo 2 (em fase de revisão pelos autores): Igor Bezerra Reis e Ariel Soares Teles. *A Method for Trademark Similarity Analysis Leveraging Generative Large Language Models*. Com planos de submissão até maio de 2025 para a revista *World Patent Information*, em específico, para a edição especial para trabalhos que utilizam IA aplicada a PI.

Referências

- Agência Sebrae de Notícias. *Em 2023, metade dos pedidos para registro de marca junto ao INPI foi feito por pequenos negócios*. 2023. <<https://agenciasebrae.com.br/cultura-empresendedora/em-2023-metade-dos-pedidos-para-registro-de-marca-junto-ao-inpi-foi-feito-por-pequenos-negocios/>>. Accessed: 2024-02-07. Citado na página 18.
- AI, U. *Unsloth: Efficient Fine-Tuning and Deployment for Large Language Models*. 2024. <<https://github.com/unslothai/unsloth>>. Accessed: 2024-11-15. Citado na página 50.
- ALSHOWAISH, H.; AL-OHALI, Y.; AL-NAFJAN, A. Trademark image similarity detection using convolutional neural network. *Applied Sciences*, v. 12, n. 3, 2022. ISSN 2076-3417. Citado 2 vezes nas páginas 17 e 38.
- BAO, Y. *et al.* Region-based cnn for logo detection. In: *Proceedings of the International Conference on Internet Multimedia Computing and Service*. New York, NY, USA: Association for Computing Machinery, 2016. (ICIMCS'16), p. 319–322. ISBN 9781450348508. Citado na página 38.
- BOGAERT, J. *et al.* Explanation sensitivity to the randomness of large language models: The case of journalistic text classification. *Traitement Automatique des Langues*, v. 64, 2023. Citado 2 vezes nas páginas 17 e 18.
- BROWN, T. *et al.* Language models are few-shot learners. In: LAROCHELLE, H. *et al.* (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>. Citado 3 vezes nas páginas 17, 27 e 31.
- CAO, J. *et al.* Unsupervised trademark retrieval method based on attention mechanism. *Sensors*, v. 21, n. 5, 2021. ISSN 1424-8220. Disponível em: <<https://www.mdpi.com/1424-8220/21/5/1894>>. Citado na página 45.
- CHANG, Y. *et al.* A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, ACM, v. 15, n. 3, p. 39:1–39:45, 2024. Citado 4 vezes nas páginas 27, 29, 30 e 33.
- CIPO. *Instituto de Propriedade Intelectual do Canadá (CIPO)*. 1976. Disponível em: <https://www.ic.gc.ca/eic/site/cipointernet-internetopic.nsf/eng/h_wr00003.html>. Citado 2 vezes nas páginas 15 e 38.
- CONTRIBUTORS, S. *Selenium: Browser Automation*. 2023. <<https://www.selenium.dev/>>. Citado na página 42.
- CUCONASU, F. *et al.* *A Tale of Trust and Accuracy: Base vs. Instruct LLMs in RAG Systems*. 2024. Disponível em: <<https://arxiv.org/abs/2406.14972>>. Citado na página 50.

- DETTMERS, T. *bitsandbytes: 8-bit Optimizers and Quantization*. 2025. Versão 0.45.3. Disponível em: <<https://pypi.org/project/bitsandbytes/>>. Citado na página 50.
- DETTMERS, T. *et al. QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. Disponível em: <<https://arxiv.org/abs/2305.14314>>. Citado na página 51.
- GALASSI, A.; LIPPI, M.; TORRONI, P. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, v. 32, n. 10, p. 4291–4308, 2021. Citado na página 31.
- Governo do Brasil. *Depósitos de patentes e marcas cresceram em 2023 no Brasil*. 2023. <<https://www.gov.br/inpi/pt-br/central-de-conteudo/noticias/depositos-de-patentes-e-marcas-cresceram-em-2023-no-brasil>>. Accessed: 2024-02-07. Citado na página 18.
- HOI, D. S. *LOGO-net Dataset*. 2015. <<http://www.mysmu.edu.sg/faculty/chhoi/logonet/index.html>>. Accessed November 20, 2023. Citado 2 vezes nas páginas 39 e 40.
- HOI, S. C. H. *et al. LOGO-Net: Large-scale Deep Logo Detection and Brand Recognition with Deep Region-based Convolutional Networks*. 2015. Citado 2 vezes nas páginas 39 e 40.
- HU, E. J. *et al. LoRA: Low-Rank Adaptation of Large Language Models*. 2021. Disponível em: <<https://arxiv.org/abs/2106.09685>>. Citado na página 51.
- HU, Y. *et al. Instructav: Instruction fine-tuning large language models for authorship verification*. In: *Proceedings of the Conference on Authorship Verification (COLM 2024)*. Singapore: [s.n.], 2024. Citado 6 vezes nas páginas 18, 27, 28, 29, 30 e 34.
- INDUSTRIAL, I. N. da P. *Portaria /INPI/PR N 70*. 2022. <<https://www.gov.br/inpi/pt-br/acao-a-informacao/dados-abertos/arquivos/documentos/diversos/plano-de-dados-abertos-bienio-2022-2024.pdf>>. Accessed January 29, 2024. Citado na página 46.
- INDUSTRIAL, I. N. da P. *Revista da Propriedade Industrial*. 2023. <<http://revistas.inpi.gov.br/rpi/>>. Citado na página 24.
- INDUSTRIAL, I. N. da P. *Dados Abertos*. 2024. <<https://www.gov.br/inpi/pt-br/acao-a-informacao/dados-abertos>>. Accessed January 29, 2024. Citado na página 46.
- INDUSTRIAL, I. N. da P. *Boletim Mensal de Propriedade Industrial – Resultados de Dezembro de 2024*. 2025. Accessed: 2025-03-23. Disponível em: <https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/publicacoes/boletim-mensal-de-pi_resultados-de-dezembro-2024.pdf>. Citado na página 17.
- INDUSTRIAL, I. N. de P. *Manual de Marcas*. 2023. Accessed October 24, 2023. Citado na página 38.
- INPI. *Boletim Mensal de PI - Resultados de Dezembro 2023*. 2023. Acesso em: 17 set. 2024. Disponível em: <https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/publicacoes/boletim-mensal-de-pi_resultados-de-dezembro-2023-1.pdf>. Citado na página 16.

INPI. *Manual de Marcas*. 2023. Acesso em: 5 maio 2023a. Disponível em: <<https://manualdemarcas.inpi.gov.br/>>. Citado 10 vezes nas páginas x, 15, 16, 22, 23, 24, 25, 26, 27 e 48.

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL. *Manual de Marcas: Análise da colidência entre sinais*. Acesso em: 7 jun. 2023b. Citado na página 16.

JOLY, A.; BUISSON, O. Logo retrieval with a contrario visual query expansion. In: *Proceedings of the 17th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2009. (MM '09), p. 581–584. Citado 3 vezes nas páginas 38, 40 e 41.

JOLY, A.; BUISSON, O. *BelgaLogos Dataset*. 2023. <<https://www-sop.inria.fr/members/Alexis.Joly/BelgaLogos/BelgaLogos.html>>. Accessed November 08, 2023. Citado 3 vezes nas páginas 38, 40 e 41.

KAMATH, U.; LIU, J.; WHITAKER, J. *Deep learning for NLP and speech recognition*. [S.l.]: Springer, 2019. v. 84. Citado na página 31.

KESIDIS, A.; KARATZAS, D. Logo and trademark recognition. In: _____. *Handbook of Document Image Processing and Recognition*. London: Springer London, 2014. p. 591–646. ISBN 978-0-85729-859-1. Disponível em: <https://doi.org/10.1007/978-0-85729-859-1_16>. Citado na página 38.

KO, K. P. *et al.* 2-gram-based Phonetic Feature Generation for Convolutional Neural Network in Assessment of Trademark Similarity. 2018. Citado 2 vezes nas páginas 34 e 36.

KOROTEEV, M. V. *BERT: A Review of Applications in Natural Language Processing and Understanding*. 2021. Citado na página 31.

KUMAR, M.; BHATIA, R.; RATTAN, D. A survey of web crawlers for information retrieval. *WIREs Data Mining and Knowledge Discovery*, v. 7, n. 6, 2017. Citado na página 39.

LAN, T. *et al.* Similar trademark image retrieval based on convolutional neural network and constraint theory. In: *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. [S.l.: s.n.], 2018. p. 1–6. Citado na página 44.

LAN, T. *et al.* Similar trademark image retrieval integrating lbp and convolutional neural network. In: ZHAO, Y.; KONG, X.; TAUBMAN, D. (Ed.). *Image and Graphics*. Cham: Springer International Publishing, 2017. p. 231–242. ISBN 978-3-319-71598-8. Citado 2 vezes nas páginas 38 e 45.

LIU, P. *et al.* Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 9, jan. 2023. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/3560815>>. Citado na página 29.

LIU, Y. *et al.* Similar trademark detection via semantic, phonetic and visual similarity information. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2021. (SIGIR '21), p. 2025–2030. ISBN 9781450380379. Disponível

em: <<https://doi.org/10.1145/3404835.3463038>>. Citado 5 vezes nas páginas 25, 35, 36, 40 e 44.

MACÍAS, W.; CERVIÑO, J. Trademark dilution and its practical effect on purchase decision. *Spanish J. Mark. ESIC*, v. 21, p. 1–13, 2017. Citado 2 vezes nas páginas 22 e 23.

MINAEE, S. *et al.* *Large Language Models: A Survey*. 2024. Disponível em: <<https://arxiv.org/abs/2402.06196>>. Citado 5 vezes nas páginas 17, 18, 27, 29 e 33.

NGUYEN, T. T. *et al.* Timely and continuous machine-learning-based classification for interactive ip traffic. *IEEE/ACM Transactions On Networking*, IEEE, v. 20, n. 6, p. 1880–1894, 2012. Citado na página 83.

NIU, Z.; ZHONG, G.; YU, H. A review on the attention mechanism of deep learning. *Neurocomputing*, v. 452, p. 48–62, 2021. ISSN 0925-2312. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S092523122100477X>>. Citado 2 vezes nas páginas x e 31.

OMPI. *Organização Mundial de Propriedade Intelectual (OMPI)*. Disponível em: <<https://www.wipo.int/>>. Citado 2 vezes nas páginas 15 e 38.

ORGANIZATION, W. I. P. *Nice Classification - WIPO - World Intellectual Property Organization*. 2024. <<https://www.wipo.int/classifications/nice/en/>>. Accessed January 24, 2024. Citado na página 44.

ORGANIZATION, W. I. P. *Vienna Classification - WIPO*. 2024. <<https://www.wipo.int/classifications/vienna/en/>>. Accessed: 2024-01-24. Citado na página 44.

Organização Mundial da Propriedade Intelectual (WIPO). *Declínio nos pedidos internacionais de marcas em 2022*. <https://www.wipo.int/pct/pt/news/2023/news_0006.html>. Accessed: 2024-02-07. Citado na página 18.

Organização Mundial da Propriedade Intelectual (WIPO). *World Intellectual Property Indicators 2018*. <https://www.wipo.int/pressroom/en/articles/2018/article_0012.html>. Accessed: 2024-02-07. Citado na página 18.

PROPERTY, N. I. of I. *Classification of products and services - INPI*. 2024. <<https://www.gov.br/inpi/pt-br/servicos/marcas/classificacao-marcas/classificacao>>. Accessed January 24, 2024. Citado na página 44.

PROPERTY, S. A. for I. *Saudi Authority for Intellectual Property*. Disponível em: <<https://www.saip.gov.sa>>. Citado na página 23.

PSYLLOS, A. P.; ANAGNOSTOPOULOS, C.-N. E.; KAYAFAS, E. Vehicle logo recognition using a sift-based enhanced matching scheme. *IEEE Trans. Intell. Transp. Syst.*, 2010. Citado na página 38.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. OpenAI, 2018. Citado na página 31.

RAFFEL, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, v. 21, n. 140, p. 1–67, 2020. Disponível em: <<http://jmlr.org/papers/v21/20-074.html>>. Citado na página 31.

- REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of EMNLP*, 2019. Citado na página 33.
- REIS, I. B. *et al.* Conflicting marks archive dataset: A dataset of conflicting marks from the brazilian intellectual property office. *Data*, v. 9, n. 2, 2024. ISSN 2306-5729. Citado 2 vezes nas páginas 47 e 84.
- ROMBERG, S. *et al.* Scalable logo recognition in real-world images. In: ACM. *ACM International Conference on Multimedia Retrieval 2011 (ICMR11)*. Trento, 2011. Also Technical Report, University of Augsburg, Institute of Computer Science, March 2011. Citado 3 vezes nas páginas 39, 40 e 41.
- ROMBERG, S. *et al.* *FlickrLogos*. 2023. <<https://www.uni-augsburg.de/en/fakultaet/fai/informatik/prof/mmc/research/datensatze/flickrlogos/>>. Accessed November 08, 2023. Citado 3 vezes nas páginas 39, 40 e 41.
- SHOWKATRAMANI, G. *et al.* Trademark image similarity search. In: STEPHANIDIS, C. (Ed.). *HCI International 2018 – Posters’ Extended Abstracts*. Cham: Springer International Publishing, 2018. p. 199–205. ISBN 978-3-319-92270-6. Citado 2 vezes nas páginas 17 e 23.
- SHOWKATRAMANI, G. *et al.* Trademark image similarity search. In: STEPHANIDIS, C. (Ed.). *HCI International 2018 – Posters’ Extended Abstracts*. Cham: Springer International Publishing, 2018. p. 199–205. ISBN 978-3-319-92270-6. Citado 2 vezes nas páginas 23 e 24.
- SMEULDERS, A. *et al.* Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 12, p. 1349–1380, 2000. Citado na página 38.
- TOUVRON, H. *et al.* *LLaMA: Open and Efficient Foundation Language Models*. 2023. Citado na página 27.
- TRAPPEY, A.; TRAPPEY, C.; LIN, S. Detecting trademark image infringement using convolutional neural networks. In: _____. [S.l.: s.n.], 2019. ISBN 9781643680200. Citado na página 44.
- TRAPPEY, A. J.; TRAPPEY, C. V.; SHIH, S. An intelligent content-based image retrieval methodology using transfer learning for digital ip protection. *Advanced Engineering Informatics*, v. 48, p. 101291, 2021. ISSN 1474-0346. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1474034621000458>>. Citado na página 44.
- TRAPPEY, C. V.; TRAPPEY, A. J. C.; LIN, S. C.-C. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. *Advanced Engineering Informatics*, v. 45, p. 101120, Aug 2020. Citado 6 vezes nas páginas 15, 16, 22, 23, 34 e 36.
- TURSUN, O.; AKER, C.; KALKAN, S. *A Large-scale Dataset and Benchmark for Similar Trademark Retrieval*. 2017. Citado 2 vezes nas páginas 39 e 40.
- TURSUN, O.; AKER, C.; KALKAN, S. *Metu Trademark Dataset*. 2023. <<https://github.com/neouyghur/METU-TRADEMARK-DATASET>>. Accessed November 08, 2023. Citado 2 vezes nas páginas 39 e 40.

- TURSUN, O. *et al.* Component-based attention for large-scale trademark retrieval. *IEEE Transactions on Information Forensics and Security*, Institute of Electrical and Electronics Engineers (IEEE), v. 17, p. 2350–2363, 2022. ISSN 1556-6021. Disponível em: <<http://dx.doi.org/10.1109/TIFS.2019.2959921>>. Citado na página 45.
- USPTO. *United States Patent and Trademark Office*. 2024. Accessed: 15 de janeiro de 2024. Disponível em: <<https://www.uspto.gov/trademarks/basics>>. Citado 2 vezes nas páginas 15 e 24.
- VASWANI, A. *et al.* Attention is all you need. In: GUYON, I. *et al.* (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citado 3 vezes nas páginas 27, 30 e 31.
- VELMURUGAN, K.; BABOO, L. Image retrieval using harris corners and histogram of oriented gradients. *Int. J. Comput. Appl.*, v. 24, p. 6–10, 2011. Citado na página 45.
- VESNIN, D.; LEVSHUN, D.; CHECHULIN, A. Trademark similarity evaluation using a combination of vit and local features. *Information*, v. 14, n. 7, 2023. ISSN 2078-2489. Citado 3 vezes nas páginas 15, 16 e 38.
- WANG, J. *et al.* Logo-2k+: A large-scale logo dataset for scalable logo classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 34, n. 04, p. 6194–6201, 2020. Citado 2 vezes nas páginas 39 e 40.
- WANG, J. *et al.* Logodet-3k: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, ACM New York, NY, v. 18, n. 1, p. 1–19, 2022. Citado 3 vezes nas páginas 39, 40 e 41.
- WANG, J. *et al.* *Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification*. 2023. <<https://github.com/msn199959/Logo-2k-plus-Dataset>>. Accessed November 08, 2023. Citado 2 vezes nas páginas 39 e 40.
- WANG, J. *et al.* *LogoDet-3K: A Large-Scale Image Dataset for Logo Detection*. 2023. <<https://github.com/Wangjing1551/LogoDet-3K-Dataset>>. Accessed November 08, 2023. Citado 3 vezes nas páginas 39, 40 e 41.
- WANG, S. *et al.* Sim-gpt: Text similarity via gpt annotated data. *arXiv preprint arXiv:2312.05603*, 2023. Disponível em: <<https://arxiv.org/abs/2312.05603>>. Citado na página 34.
- WIPO. *World Intellectual Property Indicators 2019*. [S.l.], 2019. Disponível em: <https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2019.pdf>. Citado na página 16.
- YANG, J. *et al.* Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, Association for Computing Machinery, New York, NY, USA, feb 2024. ISSN 1556-4681. Just Accepted. Disponível em: <<https://doi.org/10.1145/3649506>>. Citado na página 31.
- YANG, K. *et al.* Mentallama: Interpretable mental health analysis on social media with large language models. In: *Proceedings of the ACM Web Conference 2024 (WWW '24)*. Singapore: [s.n.], 2024. Citado 4 vezes nas páginas 18, 28, 29 e 30.

ZHANG, P.; BOULOS, M. N. K. Generative ai in medicine and healthcare: Promises, opportunities and challenges. *Future Internet*, v. 15, n. 9, 2023. ISSN 1999-5903. Citado na página 32.

ZHAO, H. *et al.* Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, ACM, v. 15, n. 2, p. 20:1–20:38, 2024. Citado 4 vezes nas páginas 27, 28, 29 e 31.

ZHAO, W. X. *et al.* *A Survey of Large Language Models*. 2023. Citado na página 32.

ZHAO, Z. *et al.* Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, p. 1–20, 2024. Citado na página 32.

Apêndices

Apêndices

Os apêndices a seguir contêm materiais complementares elaborados pelo autor, utilizados para apoiar o desenvolvimento da pesquisa, como instruções de avaliação e exemplos de prompts

Apêndice A Exemplo de Prompt Utilizado

Prompt utilizado para avaliar se duas marcas são semelhantes, com exemplos de como deve ser feita a avaliação:

Prompt para modelo explicador com exemplo de análise

Você é um especialista em avaliar similaridade de marcas. Em seguida há alguns exemplos de como você deve realizar a análise de similaridade de marcas. Por exemplo: - Comando 1: Realize uma análise da similaridade entre as marcas nominativas 'LAR PRÁTICO' (Classe Nice: NCL(11)) e 'ALTO VALE DISTRIBUIDORA' (Classe Nice: NCL(11)) e explique porque essas marcas são diferentes e não causam risco de confusão ou associação indevida entre os nomes. Considere os seguintes aspectos:

1. Aspecto fonético: Avalie a pronúncia dos nomes levando em conta a entonação, ritmo e sequência das sílabas.
2. Aspecto ideológico: Verifique se os nomes das marcas evocam ideias semelhantes, levando em consideração a classe Nice associada, mesmo que suas pronúncias sejam diferentes.
3. Distintividade: Determine se os nomes possuem elementos dominantes semelhantes.
4. Afinidade mercadológica: Analise a proximidade entre os produtos ou serviços associados às marcas, com base nas classes Nice correspondentes. - Resposta 1: Análise de diferença entre as marcas nominativas 'LAR PRÁTICO' e 'ALTO VALE DISTRIBUIDORA' (Classe Nice NCL(11))

1. Aspecto fonético: - Pronúncia: As marcas "LAR PRÁTICO" e "ALTO VALE DISTRIBUIDORA" possuem ritmos, entonação e sequência de sílabas completamente diferentes. "LAR PRÁTICO" tem uma pronúncia curta e direta, com duas palavras e um total de quatro sílabas. Já "ALTO VALE DISTRIBUIDORA" tem um nome mais longo, com sete sílabas, uma estrutura mais complexa e uma cadência diferente. Além disso, não há sons fonéticos idênticos entre os termos principais de cada marca, o que reduz significativamente qualquer risco de confusão auditiva.

2. Aspecto ideológico: - Ideia evocada: "LAR PRÁTICO" sugere praticidade e soluções voltadas para o lar, indicando um foco em produtos ou serviços relacionados a organização, manutenção ou itens domésticos. Por outro lado, "ALTO VALE DISTRIBUIDORA" remete a uma localização geográfica específica e a um serviço de distribuição, sugerindo uma

empresa que atua no fornecimento e logística de produtos. Como as ideias centrais das marcas são distintas, não há risco de os consumidores associarem um conceito ao outro.

3. Distintividade: - Elementos dominantes: As palavras-chave principais de cada marca são completamente diferentes. "LAR PRÁTICO" tem como elementos centrais "LAR" e "PRÁTICO", enquanto "ALTO VALE DISTRIBUIDORA" tem como elementos dominantes "ALTO VALE" e "DISTRIBUIDORA". Não há sobreposição significativa de termos nem palavras fonética ou semanticamente similares, o que garante uma alta distintividade entre as duas marcas.

4. Afinidade mercadológica: - Classe Nice 11: Embora ambas as marcas estejam na mesma classe Nice (11), que pode abranger produtos similares, suas propostas de valor e escopos de atuação são distintos. "LAR PRÁTICO" sugere um foco em itens para o lar, enquanto "ALTO VALE DISTRIBUIDORA" indica uma empresa de distribuição, que pode atuar fornecendo diversos produtos para diferentes setores. Como não há um direcionamento mercadológico evidente que sugira que ambas operam exatamente no mesmo nicho, a possibilidade de confusão entre os consumidores é reduzida.

Conclusão: As marcas "LAR PRÁTICO" e "ALTO VALE DISTRIBUIDORA" são suficientemente distintas para evitar qualquer risco de confusão ou associação indevida. Elas diferem significativamente em termos de fonética, ideologia e elementos dominantes, e não apresentam uma afinidade mercadológica forte o suficiente para que os consumidores presumam que pertencem a um mesmo grupo ou fornecem produtos ou serviços equivalentes. Portanto, a coexistência dessas marcas no mercado não representa um risco de confusão para os consumidores.

- Comando 2: Realize uma análise da similaridade entre as marcas nominativas 'INOVE LINGERIE' (Classe Nice NCL(11)) e 'INOVENT' (Classe Nice NCL(11) para acessórios de regulagem e segurança para aparelhos de gás) e explique porque essas marcas são diferentes e não causam risco de confusão ou associação indevida entre os nomes. Considere os seguintes aspectos:

1. Aspecto fonético: Avalie a pronúncia dos nomes levando em conta a entonação, ritmo e sequência das sílabas.

2. Aspecto ideológico: Verifique se os nomes das marcas evocam ideias semelhantes, levando em consideração a classe Nice associada, mesmo que suas pronúncias sejam diferentes.

3. Distintividade: Determine se os nomes possuem elementos dominantes semelhantes.

4. Afinidade mercadológica: Analise a proximidade entre os produtos ou serviços associados às marcas, com base nas classes Nice correspondentes. - Resposta 2: ### Análise da diferença entre as marcas 'INOVE LINGERIE' e 'INOVENT' (Classe Nice NCL(11))

1. Aspecto fonético: - Pronúncia: Embora ambas as marcas compartilhem a raiz "INOV", suas pronúncias são bastante distintas. - "INOVE LINGERIE" tem uma pronúncia fluida e suave, com quatro sílabas (I-NO-VE LIN-GE-RIE), destacando a entonação francesa na palavra "LINGERIE". - "INOVENT" possui uma pronúncia mais curta e firme, com três sílabas (I-NO-VENT), terminando com um som mais seco devido à consoante "T". - O ritmo e entonação diferentes dificultam a confusão entre os dois nomes em situações orais

ou escritas.

2. Aspecto ideológico: - "INOVE LINGERIE" remete ao setor de moda íntima, associando-se a conceitos de sofisticação, feminilidade e design elegante. - "INOVENT" sugere inovação e tecnologia industrial, uma vez que se refere a acessórios para regulagem e segurança de aparelhos de gás. - Não há sobreposição conceitual entre os dois, pois um foca no segmento de vestuário/moda, enquanto o outro está relacionado a equipamentos de segurança e regulagem de gás.

3. Distintividade: - O único elemento semelhante entre as marcas é a raiz "INOV", que pode remeter à inovação. - No entanto, o restante das palavras cria uma diferenciação clara: - "LINGERIE" indica diretamente o tipo de produto. - "ENT" em "INOVENT" sugere um sufixo mais técnico e industrial. - Essa diferença faz com que os consumidores identifiquem facilmente que se trata de segmentos distintos.

4. Afinidade mercadológica: - Apesar de ambas as marcas estarem registradas na Classe Nice 11, essa classe abrange produtos muito diferentes. - "INOVE LINGERIE" está focada em vestuário íntimo, enquanto "INOVENT" atende ao mercado de acessórios para regulagem e segurança de gás. - O público-alvo e os canais de distribuição são completamente distintos, reduzindo significativamente qualquer possibilidade de associação indevida.

Conclusão: As marcas "INOVE LINGERIE" e "INOVENT" são claramente diferentes e não apresentam risco de confusão ou associação indevida. - Diferem foneticamente, possuindo entonações e ritmos distintos. - Evocam conceitos diferentes, uma relacionada à moda íntima, e a outra a acessórios industriais. - Os elementos dominantes são distintos, com "LINGERIE" enfatizando vestuário e "VENT" remetendo a tecnologia e inovação técnica. - Não possuem afinidade mercadológica, pois atuam em nichos completamente separados dentro da Classe Nice 11.

Portanto, não há risco de os consumidores associarem as duas marcas como sendo da mesma empresa ou segmento.

Instrução: Agora, você deve avaliar a similaridade entre as seguintes marcas, considerando o contexto fornecido.

Contexto:

Realize uma análise da similaridade entre as marcas nominativas 'ROSSET LINGERIE' (Classe Nice: NCL(8) 35, para agência de importação e exportação) e 'INOVE LINGERIE' (Classe Nice: NCL(11) 35, para Comércio através de qualquer meio de lingerie, pijamas, biquíni, camisolas.) e explique porque essas marcas são diferentes e não causam risco de confusão ou associação indevida entre os nomes. Considere os seguintes aspectos:

1. Aspecto fonético: Avalie a pronúncia dos nomes levando em conta a entonação, ritmo e sequência das sílabas.
2. Aspecto ideológico: Verifique se os nomes das marcas evocam ideias semelhantes, levando em consideração a classe Nice associada, mesmo que suas pronúncias sejam diferentes.
3. Distintividade: Determine se os nomes possuem elementos dominantes semelhantes.

4. Afinidade mercadológica: Analise a proximidade entre os produtos ou serviços associados às marcas, com base nas classes Nice correspondentes.

Identifique os termos semelhantes, se houver, mas explique porque ainda há a diferença entre os nomes. Você deve gerar uma resposta com a mesma estrutura dos exemplos fornecidos.

Resposta: