Universidade Federal da Paraíba
Centro de Ciências Sociais Aplicadas
Graduação em Ciência de Dados para Negócios

# Unificação de Avaliações de Produtos em E-commerce: Uma Solução Computacional para Apoiar Decisões de Compra Online

Lívia Fernandes da Rocha

João Pessoa - PB
2026

Lívia Fernandes da Rocha

# Unificação de Avaliações de Produtos em E-commerce: Uma Solução Computacional para Apoiar Decisões de Compra Online

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência de Dados para Negócios do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Ciência de Dados para Negócios.

Orientador: Jorge Henrique Norões Viana

João Pessoa - PB
2026

**Catalogação na publicação**
**Seção de Catalogação e Classificação**

R672u Rocha, Lívia Fernandes da.
Unificação de avaliações de produtos em E-commerce: uma solução computacional para apoiar decisões de compra online / Lívia Fernandes da Rocha. - João Pessoa, 2026.
46 f. : il.

Orientação: Jorge Henrique Norões Viana.
TCC (Graduação) - UFPB/CCSA.

1. Comércio eletrônico. 2. Web scraping. 3. Integração de dados. 4. Processamento de linguagem natural. 5. Avaliações online. I. Viana, Jorge Henrique Norões. II. Título.

UFPB/CCSA CDU 004.65:658(043)

Elaborado por ANDRE DOMINGOS DA SILVA FERNANDES - CRB-15/00730

Lívia Fernandes da Rocha

# Unificação de Avaliações de Produtos em E-commerce: Uma Solução Computacional para Apoiar Decisões de Compra Online

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência de Dados para Negócios do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Ciência de Dados para Negócios.

Orientador: Jorge Henrique Norões Viana

Trabalho aprovado. João Pessoa - PB, 07 de abril de 2026:

Documento assinado digitalmente
gov.br
JORGE HENRIQUE NOROES VIANA
Data: 24/04/2026 08:20:44-0300
Verifique em https://validar.iti.gov.br

**Prof. Dr. Jorge Henrique Norões Viana**
Orientador

Documento assinado digitalmente
gov.br
ALESSIO TONY CAVALCANTI DE ALMEIDA
Data: 24/04/2026 09:40:13-0300
Verifique em https://validar.iti.gov.br

**Prof. Dr. Aléssio Tony Cavalcanti de Almeida**
Examinador

Documento assinado digitalmente
gov.br
ANTONIO VINICIUS BARROS BARBOSA
Data: 24/04/2026 08:56:28-0300
Verifique em https://validar.iti.gov.br

**Prof. Dr. Antonio Vinícius Barros Barbosa**
Examinador

João Pessoa - PB

2026

# Agradecimentos

Agradeço, primeiramente, a Deus, por me conceder força, sabedoria e perseverança ao longo de toda essa trajetória. Aos meus pais, Francisca e Josinaldo, expresso minha mais profunda gratidão, por terem feito o possível e o impossível para que o sonho da minha graduação se tornasse realidade. Todo esforço, apoio, incentivo e amor dedicados a mim foram essenciais para que eu chegasse até aqui. Esta conquista também é de vocês, pois debaixo do sol, me fizeram chegar até aqui na sombra.

Ao meu namorado, Thalles Garcia, agradeço por estar ao meu lado sempre, sendo não apenas meu parceiro de vida, mas também meu grande companheiro de estudos. Obrigada pelo apoio constante, pela paciência, pelas palavras de incentivo e, principalmente, por nunca me deixar desistir, mesmo nos momentos mais incertos.

Às minhas amigas e colegas de curso, Bianca, Lorena e Tainá, agradeço por todo o apoio, companheirismo e parceria ao longo dessa caminhada. Compartilhar os desafios, aprendizados, inseguranças e conquistas com vocês tornou essa jornada mais leve, especial e significativa.

Ao meu orientador, professor Jorge Viana, expresso minha sincera gratidão pela paciência, dedicação e pelas valiosas contribuições e ideias que enriqueceram este projeto. Agradeço também a todo o corpo docente do curso de Ciência de Dados para Negócios (CDN), ao Centro de Ciências Sociais Aplicadas (CCSA) e à Universidade Federal da Paraíba (UFPB), instituições fundamentais para a minha formação acadêmica, profissional e pessoal.

Por fim, levo comigo uma frase que sempre me marcou: “As coisas mais incríveis que podem acontecer a um ser humano, acontecerão com você, se você apenas diminuir suas expectativas.” — Phil Dunphy, Modern Family.

# Resumo

Este trabalho teve como objetivo desenvolver uma plataforma web voltada à coleta, integração e apresentação unificada de avaliações de produtos oriundos de plataformas de comércio eletrônico, com foco na Amazon e no Mercado Livre. A pesquisa partiu do problema da fragmentação das avaliações online, que dificulta a comparação entre produtos e torna o processo de decisão de compra mais demorado e disperso para o consumidor. Para enfrentar essa questão, o estudo adotou uma abordagem aplicada e experimental, estruturada em etapas de extração automatizada dos dados, tratamento e padronização das informações coletadas, integração entre plataformas e desenvolvimento de uma aplicação web para visualização comparativa dos resultados. A coleta foi realizada por meio de técnicas de web scraping, com uso das bibliotecas Selenium, BeautifulSoup e Pandas, enquanto a correspondência entre produtos equivalentes nas duas plataformas foi realizada com base em técnicas de Processamento de Linguagem Natural, utilizando TF-IDF e similaridade de cossenos. Como resultado, foi desenvolvida uma plataforma funcional capaz de consolidar avaliações e informações comparativas de 124 produtos mais vendidos da Amazon e do Mercado Livre, reunindo 16.568 avaliações coletadas. A aplicação permite visualizar, em um único ambiente, dados como preço, nota geral, quantidade de avaliações e comentários, além de identificar automaticamente a plataforma com o melhor preço e apresentar resumos gerados por inteligência artificial a partir dos comentários coletados, contribuindo para tornar o processo de busca e comparação mais simples e eficiente para o consumidor.



**Palavras-chave:** comércio eletrônico; web scraping; integração de dados; processamento de linguagem natural; avaliações online.

# Abstract

This study aimed to develop a web platform designed for the collection, integration, and unified presentation of product reviews from e-commerce platforms, focusing on Amazon and Mercado Livre. The research was motivated by the problem of fragmented online reviews, which makes product comparison more difficult and turns the consumer decision-making process into a more time-consuming and dispersed task. To address this issue, the study adopted an applied and experimental approach, structured in stages of automated data extraction, processing and standardization of the collected information, cross-platform integration, and the development of a web application for comparative visualization of results. Data collection was carried out through web scraping techniques using the Selenium, BeautifulSoup, and Pandas libraries, while the matching of equivalent products across the two platforms was performed based on Natural Language Processing techniques, using TF-IDF and cosine similarity. As a result, a functional platform was developed, capable of consolidating reviews and comparative information for 124 best-selling products from Amazon and Mercado Livre, gathering 16.568 collected reviews. The application allows users to visualize, within a single environment, data such as price, overall rating, number of reviews, and comments, as well as automatically identify the platform offering the best price and present summaries generated by artificial intelligence based on the collected comments, thereby contributing to a simpler and more efficient search and comparison process for consumers.



**Keywords:** e-commerce; web scraping; data integration; natural language processing; online reviews.

# LISTA DE FIGURAS

# LISTA DE QUADROS

# LISTA DE TABELAS

# SUMÁRIO

# 1. Introdução

O comércio eletrônico tem se consolidado nas últimas décadas como um dos pilares do consumo contemporâneo, promovendo uma profunda transformação nas relações de compra e venda. A difusão da internet e a popularização dos dispositivos móveis impulsionaram a emergência de um ambiente de consumo digital, no qual milhões de transações são realizadas diariamente em escala global. No Brasil, o e-commerce apresenta crescimento consistente nos últimos anos, com faturamento de R$ 185,7 bilhões em 2023 e projeção de R$ 204,3 bilhões em 2024 e R$ 235,5 bilhões em 2025, além da estimativa de 418,6 milhões de pedidos e 91 milhões de consumidores online em 2024, o que reforça sua relevância econômica e social no cenário atual (ABCOMM, 2025).

Nesse contexto, o processo de tomada de decisão do consumidor é fortemente influenciado por mecanismos de prova social, sobretudo pelas avaliações disponíveis em plataformas digitais. A literatura aponta que as avaliações online, conhecidas como *Electronic Word-of-Mouth* (e-WOM), representam um fator determinante para a percepção de qualidade, confiança e valor de produtos e serviços (Filieri; Mcleay, 2014). A análise dessas informações reduz as assimetrias existentes entre vendedores e compradores, funcionando como uma heurística de decisão que orienta e facilita o consumo em um ambiente marcado pela incerteza e pela abundância de alternativas.

Contudo, a fragmentação dessas avaliações em diferentes plataformas de e-commerce impõe um desafio significativo. O consumidor que deseja formar uma opinião fundamentada sobre determinado produto frequentemente se depara com a necessidade de consultar múltiplos websites (como Amazon e Mercado Livre), nos quais as informações encontram-se dispersas e apresentadas de forma heterogênea. Esse processo, além de demandar tempo e esforço, pode conduzir a decisões de compra limitadas ou enviesadas, uma vez que o indivíduo tende a basear-se apenas na parcela de dados que lhe é mais acessível.

Diante dessa problemática, surge a necessidade de desenvolver soluções que possibilitem a integração de informações oriundas de diferentes fontes, organizando-as em um formato unificado e de fácil interpretação. Neste cenário, técnicas computacionais como o web scraping e o processamento de linguagem natural apresentam-se como ferramentas adequadas para a extração, organização e padronização desses dados. O emprego dessas metodologias permite não apenas automatizar a coleta em larga escala, mas também oferecer

ao consumidor uma visão completa e precisa da reputação de produtos, auxiliando-o em seu processo decisório.

Para lidar com o excesso de informações gerado por esse grande volume de avaliações, este trabalho incorpora também um módulo de inteligência artificial, baseado no modelo *Llama 3.1 8B* da Meta, acessado via biblioteca **Hugging Face**. Esse componente é responsável por gerar automaticamente resumos textuais a partir dos comentários coletados, transformando dezenas de opiniões individuais em um relatório interpretativo de fácil leitura. Dessa forma, a plataforma não apenas reúne e compara dados dispersos, mas também oferece uma camada adicional de síntese, reduzindo a sobrecarga cognitiva do consumidor e acelerando sua decisão de compra.

O presente trabalho tem, portanto, como objetivo geral propor e desenvolver uma plataforma web voltada à coleta, integração e apresentação unificada de avaliações de produtos oriundos de sites de comércio eletrônico nacionais. A proposta busca responder à crescente demanda por soluções que reduzam a sobrecarga informacional e que promovam maior transparência no ambiente digital. O estudo combina contribuições de natureza técnica (relacionadas à arquitetura de software, metodologias de coleta e integração de dados) e de natureza social, ao favorecer o empoderamento do consumidor por meio de informações confiáveis e centralizadas.

A relevância acadêmica e prática deste estudo reside na interseção entre ciência de dados e comportamento do consumidor digital. Do ponto de vista científico, a pesquisa contribui ao demonstrar a aplicação de técnicas de extração e processamento de dados em um problema real e contemporâneo. Do ponto de vista prático, resulta em um artefato tecnológico capaz de otimizar o processo de decisão de compra, promovendo economia de tempo, aumento da confiança e melhoria na experiência de consumo online.

Logo, como principal resultado obtido, este trabalho resultou no desenvolvimento de uma plataforma web funcional, capaz de consolidar avaliações e informações comparativas dos produtos mais vendidos da Amazon e do Mercado Livre, reunindo, até o momento 16.568 avaliações coletadas de 124 produtos. A plataforma permite centralizar, em um único ambiente, dados como preço, nota geral, volume de avaliações e comentários, além de identificar automaticamente a plataforma com a melhor condição de preço para cada item analisado. Adicionalmente, o sistema passou a incorporar um recurso de inteligência artificial

para geração automática de relatórios textuais a partir dos comentários coletados, oferecendo ao usuário uma síntese interpretativa das percepções registradas pelos consumidores. Com isso, o protótipo amplia sua utilidade prática ao não apenas reunir informações dispersas, mas também organizá-las de forma amigável, contribuindo para tornar o processo de busca e comparação mais simples, ágil e eficiente para o consumidor. O protótipo encontra-se disponível para acesso em: *https://projeto-tcc-livia.onrender.com*.

Assim, esta introdução delineia o problema central a ser enfrentado e antecipa a contribuição do trabalho. Nos capítulos seguintes, serão apresentados a justificativa, os objetivos gerais e específicos, a fundamentação teórica, a metodologia adotada, a caracterização dos dados coletados e, por fim, os resultados alcançados, acompanhados de imagens da plataforma desenvolvida e de um mapa conceitual do sistema.

## 2. Justificativa

O processo de decisão de compra online é cada vez mais influenciado pela análise de avaliações de outros consumidores. No entanto, a fragmentação dessas informações em diferentes sites de e-commerce constitui uma barreira significativa. O consumidor é obrigado a realizar uma pesquisa manual exaustiva, navegando entre portais como Mercado Livre, Amazon, Shopee, Magalu, entre outros, para obter uma visão completa do produto de interesse. Esse cenário, além de ineficiente, pode levar a decisões de compra subótimas.

O projeto se justifica por sua capacidade de resolver a dispersão de avaliações de forma automatizada e eficiente. Ao centralizar dados dessas duas plataformas, a solução proposta otimiza o tempo do usuário e oferece uma perspectiva mais abrangente sobre a reputação e a qualidade dos produtos. A utilização de técnicas de web scraping se mostra viável e pertinente, dado o caráter público e acessível das informações de avaliação em sites de comércio eletrônico. A plataforma atua como uma ferramenta de empoderamento do consumidor, ao transformar a sobrecarga de dados em um recurso organizado e fácil de utilizar.

# 3. Objetivos

## 3.1. Objetivo Geral

Desenvolver e implementar uma plataforma web para a coleta, integração e apresentação unificada de avaliações de produtos de duas plataformas de e-commerce, empregando técnicas de web scraping para a extração automatizada de dados e métodos de *Processamento de Linguagem Natural (PLN)* para a unificação e análise semântica das informações coletadas.

## 3.2. Objetivos Específicos

- Identificar e selecionar as principais plataformas de e-commerce brasileiras para a coleta de avaliações, considerando sua relevância e volume de dados.
- Desenvolver e otimizar scripts de web scraping para a extração eficiente de dados dinâmicos, garantindo a adesão a boas práticas e políticas de uso dos websites.
- Projetar e implementar uma estrutura de organização e armazenamento das avaliações e metadados dos produtos.
- Aplicar técnicas de *Processamento de Linguagem Natural (PLN)*, como *TF-IDF* e similaridade de cossenos, para identificar e unificar produtos semanticamente semelhantes entre diferentes plataformas.
- Desenvolver uma interface web intuitiva e responsiva para a consulta e a visualização das avaliações agregadas, com funcionalidades de busca e filtragem.
- Implementar um mecanismo de sumarização automática dos comentários coletados, com apoio de modelos de inteligência artificial, visando facilitar a interpretação das avaliações pelos usuários.

# 4. Fundamentação Teórica

A presente seção estabelece as bases teóricas que sustentam o desenvolvimento deste trabalho. Será realizada uma imersão nos conceitos essenciais de comércio eletrônico e a psicologia do comportamento do consumidor online, seguida de uma análise técnica das metodologias de extração de dados da web (*web scraping*). Adicionalmente, serão explorados os desafios inerentes à integração de dados de fontes heterogêneas e as estratégias para sua

superação. Por fim, a seção detalhará os princípios de arquitetura de software e as tecnologias de desenvolvimento de plataformas web que orientarão a construção do protótipo.

## 4.1. Comércio Eletrônico e Comportamento do Consumidor Online

O comércio eletrônico, ou e-commerce, representa uma evolução disruptiva nas práticas comerciais, redefinindo a interação entre consumidores e empresas. Sua trajetória pode ser segmentada em fases distintas, desde as transações pioneiras baseadas em EDI (*Electronic Data Interchange*) até a era atual do comércio social e onipresente (Laudon; Traver, 2021). A fase de crescimento exponencial, a partir dos anos 2000, foi impulsionada pela massificação da internet e pelo surgimento de grandes varejistas (Albertin, 2010). Atualmente, o e-commerce 4.0 é definido pela integração com redes sociais (*social commerce*), pela predominância do acesso via dispositivos móveis (*mobile-first*) e pela personalização em massa através de algoritmos, criando uma experiência de compra fluida e multicanal (Laudon; Traver, 2021).

O ambiente digital do e-commerce é caracterizado por uma notável assimetria de informação: o vendedor detém um conhecimento sobre o produto muito superior ao do comprador. Para mitigar o risco percebido, os consumidores recorrem a heurísticas de decisão, sendo as avaliações online uma das mais poderosas. Este fenômeno, conhecido como *e-WOM* (*Electronic Word-of-Mouth*), é a versão digital da tradicional comunicação "boca a boca" e funciona como um mecanismo de prova social (Cheung; Thadani, 2012). A presença de um grande volume de avaliações, especialmente as positivas, sinaliza qualidade e confiabilidade, influenciando diretamente a intenção de compra e as taxas de conversão (Filieri; Mcleay, 2014).

Nesse sentido, as avaliações online deixam de ser apenas manifestações individuais de satisfação ou insatisfação e passam a constituir um ativo informacional relevante no processo de decisão de compra. Entretanto, quando essas informações se encontram dispersas em diferentes plataformas, o consumidor precisa empreender um esforço adicional de busca, comparação e interpretação, o que reduz a eficiência do processo decisório. Assim, compreender o papel do e-WOM e do comportamento do consumidor online fornece a base teórica para propostas que visam centralizar, integrar e tornar mais acessíveis essas informações, como a solução desenvolvida neste trabalho.

## 4.2. Web Scraping e LGPD

*Web scraping* é o processo computacional de extração automatizada de dados de websites. O processo envolve o envio de uma requisição HTTP para obter o conteúdo da página, seguido do uso de um *parser* para navegar pela estrutura do DOM (*Document Object Model*) e extrair os dados de interesse utilizando seletores específicos (Mitchell, 2018). Essa técnica permite a coleta de grandes volumes de dados de forma eficiente, viabilizando análises que seriam impraticáveis manualmente (Bhavsar Et Al., 2018). O objetivo geral deste projeto, ao empregar essa técnica para a extração sistemática de dados, está alinhado às aplicações contemporâneas do web scraping em estudos de mercado, monitoramento de reputação e análise do comportamento do consumidor.

O ecossistema utilizado: Python, oferece múltiplas ferramentas para *web scraping*. *BeautifulSoup*, por exemplo, é uma biblioteca de *parsing*[1] para analisar a estrutura de documentos HTML estáticos. *Scrapy* é um *framework*[2] completo para *crawling*[3] de larga escala. No entanto, websites modernos são predominantemente dinâmicos, utilizando JavaScript para carregar conteúdo de forma assíncrona (Mitchell, 2018). Diante deste cenário, a escolha do Selenium foi uma decisão metodológica crucial. O Selenium é uma ferramenta de automação que controla uma instância real de um navegador, permitindo que o *scraper*[4] execute e interaja com o código JavaScript da página (Mitchell, 2018). Isso viabiliza a simulação de comportamentos complexos, como rolar a página para ativar o "carregamento infinito" de comentários ou interagir com elementos em *iframes*[5], desafios técnicos que tornariam o uso de bibliotecas mais simples inviável.

A prática de web scraping demanda não apenas competência técnica, mas também rigorosa observância de princípios éticos, operacionais e jurídicos. Do ponto de vista ético, uma diretriz central é a chamada “boa vizinhança digital”, que envolve respeitar limites razoáveis de acesso (como o arquivo *robots.txt*[6]), observando as orientações expressas pelos

[1] **Parsing**: processo de análise e interpretação da estrutura de um documento ou código, transformando-o em uma representação que possa ser manipulada por programas.
[2] **Framework**: conjunto de bibliotecas, ferramentas e padrões que oferecem uma base estruturada para o desenvolvimento de aplicações.
[3] **Crawling**: processo automatizado de navegação por páginas da web, geralmente usado para indexação ou coleta de dados.
[4] **Scraper**: programa ou script responsável por extrair dados de páginas da web.
[5] **Iframe**: elemento HTML que permite a incorporação de outra página da web dentro de uma página principal.
[6] **Robots.txt**: arquivo usado por sites para informar quais áreas podem ou não ser acessadas por robôs e ferramentas de coleta automática de dados.

websites quanto à indexação e à navegação automatizada, e evitar padrões de coleta que possam comprometer o desempenho da infraestrutura dos portais acessados. Nesse sentido, a literatura recomenda cuidados como o monitoramento da frequência de requisições, a utilização de intervalos entre acessos automatizados e a limitação do volume de interações simultâneas, de modo a reduzir riscos de sobrecarga e a preservar a estabilidade dos servidores (Mitchell, 2018). Além disso, a adoção de boas práticas metodológicas em projetos de scraping também envolve delimitação clara do escopo da coleta, documentação dos procedimentos executados, transparência quanto aos tipos de dados extraídos e avaliação contínua dos riscos associados ao tratamento dessas informações.

No plano jurídico, a discussão requer atenção especial às normas de proteção de dados e ao caráter público das informações acessadas. Embora a extração de dados disponibilizados publicamente em websites seja, em geral, tratada de forma distinta da obtenção indevida de dados sigilosos ou restritos, essa atividade deve ser conduzida com cautela, sobretudo diante de possíveis limitações estabelecidas pelos Termos de Serviço das plataformas e das exigências impostas pela legislação vigente (Krotki, 2017).

No contexto brasileiro, a Lei Geral de Proteção de Dados Pessoais (LGPD) estabelece parâmetros relevantes para o tratamento de informações, especialmente quando há possibilidade de identificação de indivíduos. Por essa razão, este projeto foi concebido de maneira a operar exclusivamente sobre dados públicos de produtos e textos de avaliações disponibilizados nas páginas acessadas, sem coletar nomes, perfis, identificadores pessoais, credenciais ou qualquer informação sensível dos usuários. A coleta restringe-se a elementos necessários à análise do produto e da percepção agregada do consumidor, preservando o anonimato dos avaliadores e respeitando o princípio da minimização dos dados. Dessa forma, o projeto não busca contornar barreiras de acesso, não realiza extração de conteúdo privado e não promove tratamento de dados pessoais para fins de identificação, mantendo-se compatível com boas práticas de pesquisa, com a responsabilidade ética no uso de dados em larga escala e com os fundamentos da LGPD (Martin, 2015).

## 4.3. Integração dos Dados

A integração de dados visa combinar informações de fontes heterogêneas para criar uma visão unificada e consistente (Halevy; Rajaraman; Ordonez, 2009). O desafio central neste processo é a resolução de entidades (*entity resolution*), que consiste em identificar e

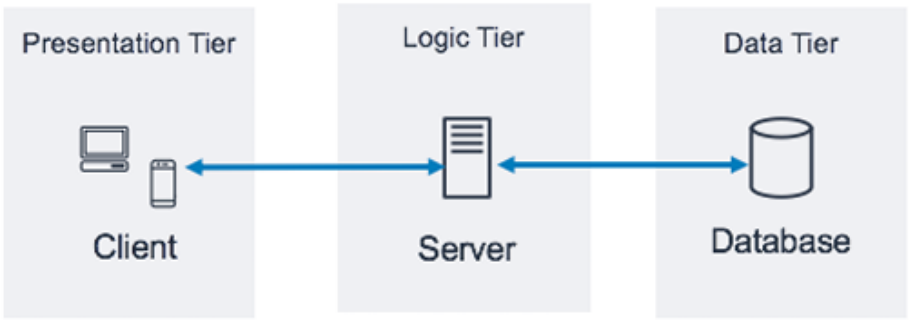
agrupar registros que se referem à mesma entidade do mundo real, apesar de possuírem representações textuais diferentes (Halevy; Rajaraman; Ordonez, 2009). Para superar essa barreira, foram aplicadas técnicas de Processamento de Linguagem Natural (PLN). Especificamente, o modelo TF-IDF (Term Frequency-Inverse Document Frequency) foi utilizado para converter os títulos dos produtos em vetores numéricos. Em seguida, a similaridade de cossenos (*cosine similarity*) foi aplicada para calcular a proximidade entre esses vetores, permitindo a identificação e o agrupamento de produtos correspondentes com alta precisão (Manning; Raghavan; Schütze, 2008).

A arquitetura de armazenamento é um pilar para qualquer sistema de integração. A escolha de um modelo de dados define como as informações heterogêneas serão harmonizadas. Enquanto bancos de dados relacionais (SQL) impõem um esquema rígido (Elmasri; Navathe, 2016), bancos de dados não-relacionais (NoSQL) oferecem flexibilidade esquemática, sendo mais adequados para dados semiestruturados como os de *web scraping* (Sadalage; Fowler, 2012). Neste projeto, um dos objetivos específicos foi o de projetar um modelo para o armazenamento estruturado. A biblioteca Pandas foi utilizada para organizar os dados em uma estrutura tabular (DataFrame), que serve como um modelo lógico limpo e de fácil manipulação, constituindo a base empírica para a plataforma de visualização. Além disso, estabelece uma estrutura de dados previamente organizada para futura implementação em banco de dados, o que amplia o potencial de escalabilidade, robustez e manutenção da aplicação.

## 4.4 Desenvolvimento de Plataformas Web

A arquitetura de software de um sistema web moderno é comumente estruturada sob o paradigma Cliente-Servidor e organizada em um padrão de três camadas (*3-tier architecture*) para promover a separação de responsabilidades (Bass; Clements; Kazman, 2012), proporcionando modularidade, escalabilidade e facilidade de manutenção. A Figura 1 ilustra a aplicação em três níveis principais: apresentação, lógica de aplicação e dados.

**Figura 1:** Padrão de uma arquitetura em três camadas (three-tier architecture).




***Fonte:*** *Amazon Web Services (2021).*

A Camada de Apresentação (Frontend) é responsável pela interface com o usuário e pela exibição das informações. A Camada de Lógica de Aplicação (Backend) processa as requisições, aplica as regras de negócio e coordena o fluxo de dados do sistema. Já a Camada de Dados é responsável pela persistência e organização das informações (Richards, 2015). No contexto deste projeto, essa arquitetura permite separar de forma clara a interface de visualização, os scripts de processamento e coleta, e a base consolidada de dados, contribuindo para a clareza estrutural e para a manutenção evolutiva do sistema (Bass; Clements; Kazman, 2012).

A seleção das tecnologias foi guiada pelos requisitos de um projeto intensivo em dados, que demanda integração entre coleta automatizada, processamento textual e disponibilização web. Para o backend, a linguagem Python foi adotada em virtude de seu amplo ecossistema de bibliotecas voltadas à automação, análise de dados e desenvolvimento de aplicações (Ramalho, 2022). Para a camada de desenvolvimento web, optou-se pelo framework Flask, que oferece uma estrutura leve, flexível e adequada à construção de aplicações sob medida, permitindo maior controle sobre rotas, templates e integração com os módulos de processamento do sistema. Na camada de apresentação, utilizaram-se tecnologias padrão da web, como HTML, CSS e JavaScript, para estruturar uma interface capaz de permitir busca, filtragem e visualização comparativa das avaliações agregadas (Groner, 2018).

Para a disponibilização pública do protótipo, a aplicação foi publicada na plataforma Render, um serviço de hospedagem em nuvem que viabiliza o deploy de aplicações web de forma prática e compatível com projetos desenvolvidos em Python. Essa combinação

tecnológica mostrou-se adequada para integrar as diferentes etapas do sistema em um protótipo funcional e acessível ao usuário final.

## 4.5 As Plataformas de Comércio Eletrônico

A escolha das plataformas analisadas foi uma etapa importante deste trabalho, pois influencia diretamente a quantidade, diversidade e comparabilidade dos dados coletados. Neste projeto, optou-se por delimitar o estudo à Amazon e ao Mercado Livre, por serem plataformas de grande relevância no comércio eletrônico e por concentrarem volume expressivo de produtos, ofertas e avaliações de consumidores. Essa escolha também é coerente com o cenário atual do setor, já que o e-commerce brasileiro segue em expansão e mantém forte adesão dos consumidores no país.

Além disso, a Amazon e o Mercado Livre ocupam posições de destaque no ambiente digital brasileiro. Segundo o ranking da Similarweb para a categoria “E-commerce e compras” no Brasil, referente a fevereiro de 2026, o Mercado Livre aparece em primeiro lugar entre os sites mais visitados do setor, enquanto a Amazon Brasil ocupa a segunda posição (SIMILARWEB, 2026).

Em termos de participação de mercado, essas plataformas também ocupam posições centrais no comércio eletrônico brasileiro. De acordo com o guia *Brazil – eCommerce*, publicado pelo U.S. Commercial Service em 2025, o Mercado Livre detém aproximadamente 35% do market share do setor no país, enquanto a Amazon Brasil concentra cerca de 20%, o que reforça a relevância competitiva de ambas no ambiente digital nacional. No caso da Amazon, sua presença também se projeta para além do próprio marketplace, uma vez que a mesma aparece em serviços de comparação e apoio à compra, como o Buscapé, o que amplia sua visibilidade no ecossistema digital de consumo.

Outro ponto que justifica essa escolha é que os sites escolhidos funcionam como grandes marketplaces, com ampla variedade de categorias e alto volume de interações entre consumidores e vendedores. No caso da Amazon Brasil, a própria plataforma informa operar com mais de 30 categorias de produtos. Já o Mercado Livre se apresenta oficialmente como o maior marketplace da América Latina. Essas características tornam ambas especialmente adequadas para um projeto que depende da coleta de dados públicos, da comparação entre produtos e da análise de avaliações em larga escala.

Dessa forma, a definição de Amazon e Mercado Livre como foco do estudo não foi aleatória, mas sim uma escolha metodológica alinhada aos objetivos do projeto. Como o propósito do trabalho é centralizar e comparar avaliações de produtos dispersas em diferentes marketplaces, faz sentido concentrar a análise em plataformas que possuem grande visibilidade, diversidade de anúncios e sistemas consolidados de reputação por notas e comentários. Assim, o recorte adotado fortalece a aplicabilidade da solução desenvolvida e garante um contexto empírico relevante para a proposta do trabalho.

# 5. Metodologia

O presente trabalho adotou uma metodologia de caráter aplicado e experimental, fundamentada na construção e validação de um pipeline completo de aquisição, integração, enriquecimento e disponibilização de dados oriundos de duas plataformas de comércio eletrônico: Amazon e Mercado Livre. A estratégia metodológica foi estruturada de forma incremental, partindo da coleta automatizada de dados brutos, passando por etapas de padronização e integração interplataforma, até a consolidação em uma aplicação web funcional capaz de operacionalizar a comparação de avaliações e preços entre diferentes marketplaces.

A coleta de dados foi implementada por meio de técnicas de web scraping aplicadas a ambientes de páginas dinâmicas, utilizando automação de navegador com Selenium para simular o comportamento de um usuário real (como clicar em botões e rolar a página) e processamento estrutural do HTML via BeautifulSoup. A escolha dessa abordagem decorre da necessidade de interação com conteúdos carregados assincronamente, elementos renderizados por JavaScript e estruturas encapsuladas em iframes, características comuns em plataformas de e-commerce contemporâneas. A biblioteca Pandas foi empregada para a organização e a manipulação tabular dos dados extraídos, facilitando a exportação para o formato CSV.

Visando o cumprimento da Lei Geral de Proteção de Dados (LGPD), a arquitetura de coleta foi projetada para trabalhar exclusivamente com dados públicos disponíveis nas páginas de produtos. Dessa forma, apenas informações relacionadas aos produtos e aos textos das avaliações são coletadas, sem o armazenamento de qualquer identificador pessoal dos usuários.

A extração contempla informações como nome do produto, preço, nota média, número total de avaliações, características técnicas obtidas a partir de tabelas estruturadas da página e variações comerciais disponíveis, como voltagem, cor e capacidade. Por fim, o scraper foi desenvolvido utilizando seletores HTML alternativos como mecanismo de contingência. Essa estratégia permite que o processo de coleta continue funcionando mesmo em situações de pequenas alterações na estrutura ou no layout das páginas da plataforma. A lógica de extração é sutilmente diferente entre as plataformas, conforme detalhado nas próximas seções.

## 5.1. Detalhamento do scraping da Amazon

A coleta de dados na plataforma Amazon foi desenvolvida de forma a manter a estabilidade do processo mesmo diante da estrutura dinâmica da página. Para melhorar o desempenho da navegação automatizada, o sistema utiliza o armazenamento local de cookies de sessão, permitindo reutilizar uma autenticação previamente realizada e reduzindo a necessidade de novos logins durante o processo de coleta.

O scraper desenvolvido para a plataforma Amazon foi projetado com mecanismos adicionais para aumentar a estabilidade do processo de coleta. Devido à natureza dinâmica da plataforma e às restrições de limite de requisições (rate limiting), a navegação automatizada foi configurada para operar com intervalos de tempo aleatórios entre as requisições e com reinicializações periódicas da sessão do navegador, reduzindo o risco de bloqueios e garantindo maior continuidade durante a extração dos dados.

A navegação automatizada percorre sobre as subcategorias da seção de 'Mais Vendidos', filtrando e ignorando departamentos de produtos exclusivamente digitais para focar em bens de consumo físicos. Para cada item identificado, a rotina de extração capturou metadados básicos (nome, preço à vista, nota média e volume de avaliações) e foi aprimorada para extrair especificações técnicas e variações comerciais (como voltagem, cor e capacidade) a partir de tabelas estruturadas e elementos de seleção dinâmica presentes na página do produto.

A função da coleta de avaliações (**extrair_comentarios_amazon**) utiliza as bibliotecas Selenium para interação e navegação, e BeautifulSoup para o processamento da estrutura DOM (Document Object Model).

Na Amazon, a extração das avaliações é realizada diretamente a partir da estrutura HTML da página do produto. O script localiza os comentários utilizando prioritariamente elementos identificados pelo atributo *data-hook='review'*. Como mecanismo adicional de segurança, também são considerados elementos cujos identificadores de divisão (div) possuem o prefixo *customer_review-*, o que torna o processo de coleta mais robusto caso haja pequenas alterações na estrutura da página. Durante a coleta, é realizada uma verificação da origem da avaliação, garantindo que apenas comentários provenientes do Brasil sejam considerados na análise. Essa filtragem é importante para manter a consistência dos dados, evitando a inclusão de avaliações de outros mercados que possam apresentar diferenças de idioma, contexto ou comportamento de consumo.

Após essa etapa, a data e o país associados à avaliação passam por um processo de padronização realizado pela função auxiliar **formatar_data_e_pais_amazon**. Essa função utiliza expressões regulares (RegEx) para identificar e organizar corretamente o formato das datas presentes nas avaliações. Os demais campos das avaliações, como o texto do comentário e a nota atribuída ao produto, são obtidos diretamente por meio de seletores CSS aplicados aos elementos correspondentes na página. O processo de coleta continua até atingir o limite de comentários definido para cada produto (usualmente 10 comentários), momento em que a extração é interrompida automaticamente e o script prossegue para o próximo item. Este processo está resumido no pseudocódigo apresentado na Figura 2.

**Figura 2:** Pseudocódigo da Extração de Comentários na Amazon (scraper3.py)

```
FUNÇÃO extrair_comentarios_amazon(driver, asin, max_comentarios):
    ACESSAR url_de_avaliacoes_do_produto_com_asin

    TENTAR:
        APLICAR_FILTRO_NA_INTERFACE_PARA("Brasil")
    IGNORAR_ERROS

    soup = PROCESSAR_HTML(driver.page_source)
    blocos_de_comentario =
ENCONTRAR_TODOS_ELEMENTOS_DE_AVALIACAO(soup)
    comentarios_validados = []

    PARA CADA bloco EM blocos_de_comentario:
        data_e_local_str = EXTRAIR_ELEMENTO_DE_DATA(bloco)

    # Filtro de Homogeneidade Territorial
        SE NÃO contem_palavra_chave("Brasil", data_e_local_str):
            CONTINUAR_PROXIMO_LOOP # Ignora comentários de outros países

        nota = EXTRAIR_NOTA_EM_ESTRELAS(bloco)
```

```
        texto_comentario = EXTRAIR_TEXTO_DA_AVALIACAO(bloco)
        data_formatada = FORMATAR_DATA_PADRAO_BR(data_e_local_str)

        ADICIONAR AO ARRAY comentarios_validados:
            "ASIN": asin
            "País": "Brasil"
            "Data Comentário": data_formatada
            "Nota Comentário": nota
            "Comentário": texto_comentario

        SE QUANTIDADE_DE_ITENS(comentarios_validados) >= max_comentarios:
            PARAR_LOOP

    RETORNAR comentarios_validados
```

***Fonte:*** *Código de Scraping da Amazon*

## 5.2. Detalhamento do scraping do Mercado Livre

A coleta de dados no Mercado Livre foi desenvolvida com o objetivo de manter a eficiência do tempo, e evitar sobrecarga de memória durante a execução do scraping. Embora as avaliações da plataforma sejam públicas, o script realiza o carregamento prévio de cookies de sessão, o que contribui para reduzir o tempo de resposta das páginas e tornar a navegação automatizada mais estável.

O scraper desenvolvido realiza a navegação na seção de produtos “Mais Vendidos”, percorrendo as subcategorias e extraindo informações estruturais e comportamentais dos produtos. Os comentários são coletados após interação automatizada com a interface de avaliações, incluindo acionamento de botões de expansão e rolagem da página para carregamento de conteúdo dinâmico. Cada comentário coletado é armazenado individualmente como uma unidade de análise, contendo o texto da avaliação, a nota atribuída pelo consumidor e a data da avaliação já padronizada.

Para garantir a estabilidade do processo de raspagem e evitar que o estado da página de listagem seja perdido durante a navegação, cada produto é aberto em uma nova aba do navegador, de forma a garantir a continuidade do processo de coleta de dados em larga escala, impedindo assim que eventuais falhas de carregamento em uma página específica interrompam todo o processo de coleta. A partir da URL de cada produto, o identificador único do item (ASIN, no caso da Amazon, ou MLB no Mercado Livre) é extraído utilizando expressões regulares (RegEx). Em seguida, o script coleta os principais metadados do produto, como nome, ID derivado da URL, preço, nota média e número total de avaliações,

além de mapear características do produto disponíveis na página, como cor, voltagem e outras especificações técnicas, além dos comentários individuais.

Uma das principais dificuldades na coleta de dados dessa plataforma está relacionada à renderização dinâmica das avaliações. Para acessar os comentários, o script automatiza a interação com elementos da interface, como o botão “Ver todas as opiniões”, simulando o comportamento de um usuário real. Após essa etapa, o algoritmo verifica a estrutura do Document Object Model (DOM) da página para identificar como os comentários foram carregados. Em alguns casos, os comentários são exibidos dentro de um elemento iframe, exigindo a mudança de contexto do driver para acessar o conteúdo. Em outros casos, as avaliações são carregadas diretamente na página principal. Para garantir que todos os comentários disponíveis sejam carregados, o script realiza múltiplas simulações de rolagem da página, permitindo o carregamento assíncrono dos elementos.

Ao final desse processo, são coletados até 10 comentários por produto, incluindo a identificação da nota atribuída pelos usuários por meio das estrelas exibidas na avaliação. Após a extração das informações, a aba do produto é fechada e o script retorna automaticamente para a página principal de listagem, continuando o processo de coleta para os demais itens. Posteriormente, todas as informações coletadas são organizadas em uma estrutura de dados padronizada. Essa organização permite a exportação dos dados para DataFrames da biblioteca Pandas, facilitando a etapa de processamento e possibilitando a integração, mais tarde, com os dados obtidos da plataforma Amazon. O processo completo para o Mercado Livre está descrito no pseudocódigo da Figura 3.

É importante destacar uma diferença estrutural entre as duas plataformas no que diz respeito à seleção dos comentários exibidos. Enquanto a Amazon tende a mostrar, por padrão, as avaliações mais recentes (com predomínio de datas de 2024 e 2025), o Mercado Livre prioriza o que seu algoritmo classifica como “comentários mais relevantes”. Essa relevância nem sempre corresponde à recência; durante a coleta, observaram-se avaliações de 2021 e 2023 sendo posicionadas no topo da lista, enquanto comentários mais novos ficavam relegados a páginas secundárias ou exigiam filtros manuais para serem acessados. Esse comportamento impacta diretamente a comparabilidade entre as plataformas, pois os conjuntos de comentários extraídos podem representar períodos distintos da vida do produto. As implicações dessa assimetria são discutidas nas considerações finais.

**Figura 3**: Pseudocódigo da Extração de Comentários no Mercado Livre (scraperml2.py)

```
FUNÇÃO extrair_dados_e_comentarios_ml(link_produto, max_comentarios=10):
    EXECUTAR_SCRIPT_JS("window.open(link_produto)")
    MUDAR_FOCO_PARA_NOVA_ABA()

    asin = EXTRAIR_ID_DA_URL_VIA_REGEX(link_produto)
    preco, nota_geral = EXTRAIR_METADADOS_BASICOS()
    caracteristicas = EXTRAIR_VARIACOES_DO_PRODUTO()

    SIMULAR_CLIQUE("Ver todas as opiniões")

    SE EXISTIR_ELEMENTO("iframe" CONTENDO "reviews"):
        MUDAR_FOCO_PARA_IFRAME(iframe)

    PARA i DE 1 ATÉ 5:
        RODAR_PAGINA_PARA_BAIXO()
        ESPERAR_CARREGAMENTO_ASSINCRONO()

    comentarios_coletados = []
    elementos_comentario =
ENCONTRAR_TODOS_OS_ARTIGOS_DE_COMENTARIO()

    PARA CADA comentario_el EM elementos_comentario:
        texto = EXTRAIR_TEXTO_DO_COMENTARIO(comentario_el)
        nota =
CALCULAR_ESTRELAS_PREENCHIDAS_E_FRACIONADAS(comentario_el)
        data = EXTRAIR_E_FORMATAR_DATA(comentario_el)

        ADICIONAR AO ARRAY comentarios_coletados:
            "Comentário": texto
            "Nota": nota
            "Data Comentário": data

        SE QUANTIDADE_DE_ITENS(comentarios_coletados) >= max_comentarios:
            PARAR_LOOP

    RETORNAR_FOCO_PARA_CONTEUDO_PRINCIPAL()
    FECHAR_ABA_ATUAL()
    MUDAR_FOCO_PARA_ABA_DE_LISTAGEM()

    RETORNAR dados_consolidados
```

***Fonte:*** *Código de Scraping do Mercado Livre*

## 5.3. Integração dos dados, vetorização e similaridade

Após a conclusão da etapa de extração automatizada, os dados brutos passaram por uma fase de organização e tratamento por meio da biblioteca Pandas, sendo selecionados os arquivos mais recentes gerados pelo processo de coleta. Para permitir a comparação entre produtos das duas plataformas, os títulos dos produtos foram submetidos a um processo de pré-processamento textual. Essa etapa incluiu a conversão de todos os caracteres para letras

minúsculas, a remoção de acentuação com auxílio da biblioteca unidecode, a eliminação de caracteres especiais por meio de expressões regulares (RegEx) e a limitação do tamanho das *strings* aos primeiros 60 caracteres. Essa estratégia permite concentrar a análise nos termos mais relevantes dos títulos dos produtos.

A integração entre os dados da Amazon e do Mercado Livre foi realizada por meio de um processo de Resolução de Entidades *(Entity Resolution)* baseado em técnicas de Processamento de Linguagem Natural (PLN). Para reduzir o custo computacional e minimizar a ocorrência de falsos positivos, foi realizado previamente um mapeamento entre categorias equivalentes das duas plataformas. Dessa forma, as comparações são realizadas apenas entre categorias inter-relacionadas, como, por exemplo, a categoria "Beleza" da Amazon com "Beleza e Cuidado Pessoal" do Mercado Livre. O quadro completo de compatibilidade entre categorias utilizado nesta etapa encontra-se apresentado no Apêndice A.

Quanto ao conjunto de produtos já filtrados por categoria, foi aplicado o método de vetorização TF-IDF (Term Frequency-Inverse Document Frequency). Essa técnica converteu os títulos limpos dos produtos em vetores numéricos, ponderando a relevância de cada termo com base em sua frequência e raridade. A partir desses vetores, o grau de similaridade entre os produtos foi calculado utilizando a Similaridade de Cossenos, onde foi definido um limiar mínimo de corte (*threshold*) de 0,70 (70%) de similaridade para considerar que dois produtos representam o mesmo item nas diferentes plataformas.

Adicionalmente, foi realizado um teste complementar com um método alternativo de similaridade, baseado na extração de atributos presentes nos nomes dos produtos e na aplicação de pesos diferenciados a esses atributos para fins de comparação. Entretanto, os resultados obtidos não apresentaram desempenho satisfatório quando comparados ao método principal baseado em TF-IDF e similaridade de cossenos. Por essa razão, esse procedimento não foi incorporado à solução final do sistema, sendo apresentado em Apêndice B.

A partir da correspondência entre produtos, foi criada uma base consolidada (Master Data). Nessa base, foi implementado um cálculo de reputação unificada do produto. Em vez de utilizar uma média simples das avaliações, foi adotada uma nota única ponderada, na qual o peso de cada avaliação é determinado pelo volume total de avaliações da respectiva plataforma, conforme apresentado na fórmula a seguir:

$$NotaFinal = \frac{(Nota_A \times Qtd_A) + (Nota_M \times Qtd_M)}{Qtd_A + Qtd_M}$$

Essa formulação matemática permite que a nota final represente de maneira mais adequada a avaliação geral dos consumidores. Ao utilizar o número de avaliações como peso no cálculo, o modelo reduz o impacto de produtos com poucas avaliações e atribui maior relevância àqueles que possuem um volume maior de feedback dos usuários.

Por fim, com o objetivo de manter as informações do sistema atualizadas sem a necessidade de executar novamente todo o processo de comparação baseado em técnicas de Processamento de Linguagem Natural, foi desenvolvido um módulo adicional de atualização automática. Esse script secundário processa periodicamente os arquivos mais recentes gerados pelos scrapers das plataformas. Durante essa etapa, o sistema identifica a última ocorrência de cada produto a partir de seus identificadores únicos (ASIN, no caso da Amazon, e ID do produto no Mercado Livre) e atualiza os preços diretamente na base consolidada (Master Data). Dessa forma, o protótipo consegue acompanhar as variações de preço das plataformas ao longo do tempo, mantendo os dados alinhados com a dinâmica do comércio eletrônico.

## 5.4. Desenvolvimento do aplicativo

A etapa de disponibilização dos dados foi implementada por meio de uma aplicação web desenvolvida em Python com o framework Flask, adotado como camada intermediária entre os arquivos consolidados gerados nas etapas anteriores e a interface acessada pelo usuário. A escolha do Flask decorreu de sua leveza estrutural e de sua flexibilidade para a construção de aplicações orientadas por rotas, templates e funções de backend customizadas. No aplicativo desenvolvido, o Flask foi utilizado em conjunto com **render_template**, para a renderização das páginas HTML, e com **request**, para a captura dos parâmetros de consulta enviados pela interface, possibilitando a construção de um fluxo interativo de busca, filtragem e detalhamento de produtos.

No backend, a biblioteca Pandas foi empregada como principal mecanismo de manipulação dos dados consumidos pela aplicação. A base comparativa principal, os arquivos amostrais da Amazon e do Mercado Livre e os arquivos completos utilizados para recuperação de comentários são carregados por funções auxiliares centralizadas no módulo

**utils.loaders**, responsável por abstrair o acesso às diferentes estruturas derivadas do pipeline de scraping e matching. Essa camada inclui funções como **carregar_comparacao_master_com_precos**, **atualizar_cache_comentarios** e **buscar_comentarios_cache_por_produto**, permitindo que a aplicação opere sobre dados previamente tratados e organizados. Como estratégia de desempenho, parte dessas estruturas é mantida em memória durante a execução, com controle de atualização periódica da base comparativa e do cache de comentários.

Além das bibliotecas centrais de aplicação e manipulação dos dados, foram utilizadas bibliotecas auxiliares voltadas ao tratamento textual, ao controle de execução e à integração com serviços externos. O módulo **re** foi empregado para normalização textual e extração de padrões; os módulos **os** e **time**, para controle de ambiente e temporização; a biblioteca **requests**, para chamadas auxiliares; e a biblioteca **Hugging Face**, que foi utilizada na implementação do modelo de IA da Meta o *Llama 3.1 8B*, para geração automática de resumos textuais a partir dos comentários coletados, onde transforma um conjunto de avaliações individuais em um resumo interpretativo mais acessível ao usuário. Essa composição de bibliotecas reflete uma arquitetura de aplicação centrada em processamento leve, atualização incremental, reaproveitamento de estruturas em memória e enriquecimento interpretativo dos dados exibidos ao usuário.

Do ponto de vista funcional, a aplicação foi organizada em rotas específicas. Na rota de busca, os parâmetros recebidos da interface são utilizados para filtrar os registros da base consolidada, sendo as categorias montadas a partir da base amostral da Amazon. Nessa mesma etapa, o backend realiza a normalização de identificadores, o preenchimento de preços ausentes por meio de funções auxiliares e o cálculo automático da melhor plataforma com base no menor preço encontrado. Dessa forma, a aplicação não se limita à exibição estática dos dados, mas incorpora regras de negócio executadas em tempo de consulta.

Na camada de apresentação, utilizaram-se templates HTML integrados ao mecanismo de renderização do Flask, complementados por CSS e JavaScript para estruturação visual, estilização e interação com o usuário. Essa combinação possibilitou a construção de páginas distintas para a tela inicial, a interface de busca e a página de detalhe do produto, preservando a separação entre lógica de aplicação e apresentação. A interface foi, portanto, desenvolvida segundo o modelo cliente-servidor, no qual o navegador atua como cliente responsável pela exibição e submissão das consultas, enquanto o backend em Flask processa as requisições,

acessa os dados e devolve as páginas renderizadas com os resultados já tratados. Essa organização está de acordo com a arquitetura em três camadas discutida na fundamentação teórica.

Por fim, a aplicação foi publicada na plataforma Render, um ambiente de hospedagem em nuvem compatível com aplicações Python, o que permitiu a disponibilização externa do protótipo. Dessa forma, o desenvolvimento do aplicativo não constituiu apenas uma etapa de visualização, mas a implementação de uma camada computacional responsável por integrar arquivos processados, regras de negócio, renderização web e mecanismos de atualização de dados em um sistema funcional orientado à comparação interplataforma de produtos.

# 6. Análise e Caracterização dos Dados Coletados

Os dados obtidos por meio do processo de web scraping representam um conjunto estruturado e multifacetado de informações essenciais para a construção da plataforma. Os datasets resultantes da coleta em plataformas como Amazon e Mercado Livre são compostos por variáveis quantitativas e qualitativas, permitindo diferentes níveis de análise, proporcionando assim uma visão da dimensão e das características da base construída ao longo do projeto.

A coleta de dados em ambos os sites seguiu um fluxo semelhante, mas com implementações distintas que processem as especificidades de cada plataforma. O Quadro *1* a seguir ilustra as principais semelhanças e diferenças no processo de extração de dados entre Amazon e Mercado Livre.

**Quadro 1:** Comparativo das Metodologias de Web Scraping.

| **Aspecto da Coleta** | **Amazon (Mais Vendidos)** | **Mercado Livre (Mais Vendidos)** |
|---|---|---|
| **Acesso** | Requer autenticação por login, com cookies persistidos. | Igualmente necessitou de autenticação por login, com cookies persistidos. |
| **Identificador Único** | ASIN (Amazon Standard Identification Number). | ID numérico (ex: MLB15578949). |

| | | |
|---|---|---|
| **Extração de Produtos** | Extração de ASINs da página de bestsellers. | Extração de IDs e links da página de mais vendidos. |
| **Carregamento de Comentários** | Rolagem da página para carregar conteúdo incremental. | Clicar no botão ("Mostrar todas as opiniões") e simular rolagem em um iframe. |
| **Campos Coletados** | Posição Global, Posição Categoria, Subcategoria, ASIN, Nome, Preço à vista, Nota Geral, Qtd. Avaliações, Características, País, Data Comentário, Nota, Comentário, Link. | Posição Global, Posição Categoria, Subcategoria, ASIN(ID), Nome, Preço à vista, Nota Geral, Qtd. Avaliações, Características, País, Data Comentário, Nota, Comentário, Link. |

***Fonte:*** *Bases geradas após scrapings*

Quanto à estrutura dos dados coletados, optou-se por uma padronização entre as duas plataformas, o que favoreceu a integração posterior. Em ambos os casos, foram identificadas colunas equivalentes para identificador do produto (ASIN), categoria, quantidade de avaliações, nota geral e comentário. Essa proximidade estrutural foi essencial para viabilizar o processo de padronização, vetorização e matching entre os produtos das duas plataformas. O Quadro *2* a seguir demonstra os principais resultados quantitativos observados nas bases de dados após a extração completa.

**Tabela 1 :** Resumo geral dos produtos e avaliações em “Mais Vendidos”.

| **Plataforma** | **Total de Produtos** | **Total de Categorias** | **Total de Avaliações** | **Média de Avaliações por Produto** | **Total Comentários Coletados** | **Média das Notas** |
|---|---|---|---|---|---|---|
| Amazon | 1.216 | 25 | 151.403 | 125 | 11.895 | 4,65 |
| Mercado Livre | 460 | 28 | 9.768.302 | 21.235 | 4.673 | 4,79 |

***Fonte:*** *Bases geradas após scrapings*

Os dados apresentados evidenciam contrastes relevantes entre as plataformas quando considerado o mesmo universo de produtos. No caso da Amazon, a extração resultou em um

total de 1.216 produtos, distribuídos em 25 subcategorias. A base também registrou 151.403 avaliações agregadas, considerando o volume de avaliações informado nas páginas dos produtos, com uma média de 125 avaliações por produto. Além disso, foram coletados 11.895 comentários individuais, o que demonstra um volume expressivo de dados textuais disponíveis para análise. A média das notas gerais dos produtos extraídos na plataforma foi de 4,65 (em uma escala de 1 a 5), indicando, de modo geral, uma avaliação positiva dos itens coletados.

Para o Mercado Livre, a coleta resultou em 460 produtos, distribuídos em 28 subcategorias. Em termos de volume de avaliações agregadas, a plataforma apresentou 9.768.302 avaliações, com uma média de 21.235 avaliações por produto, valor significativamente superior ao observado na Amazon. Foram coletados ainda 4.673 comentários individuais, e a média das notas gerais dos produtos foi de 4,79. Esses resultados indicam que, embora a base do Mercado Livre tenha apresentado menor número de produtos e comentários coletados, os itens extraídos concentram, em média, um volume muito maior de avaliações registradas na plataforma.

De modo geral, os resultados evidenciam que as duas bases apresentam perfis distintos e complementares. A Amazon contribuiu com maior diversidade de produtos e maior volume de comentários textuais coletados, enquanto o Mercado Livre se destacou pelo elevado número de avaliações agregadas por produto. Essa diferença pode estar associada tanto às características estruturais de cada plataforma quanto à forma como as avaliações são exibidas e disponibilizadas ao público. Em conjunto, essas bases forneceram suporte consistente para a etapa de comparação interplataforma, permitindo trabalhar simultaneamente com variáveis quantitativas, como nota média e quantidade de avaliações, e com conteúdo textual oriundo dos comentários dos consumidores.

# 7. Resultados

## 7.1. Percurso Metodológico e Implementação do Protótipo

O desenvolvimento do projeto avançou da etapa de planejamento e modelagem conceitual, para a implementação prática de um protótipo funcional. Após a consolidação das etapas de revisão bibliográfica, definição do escopo e desenvolvimento dos scrapers para Amazon e Mercado Livre, foi possível estruturar a etapa de integração dos dados e, em

seguida, construir a aplicação web responsável por disponibilizar os resultados ao usuário final. Dessa forma, o trabalho deixou de se limitar à coleta automatizada e passou a incorporar uma interface de consulta e visualização comparativa das informações extraídas.

A aplicação desenvolvida foi implementada com o framework Flask, em linguagem Python, mantendo a separação entre a camada de processamento dos dados e a camada de apresentação. O sistema foi estruturado para carregar a base consolidada de comparação entre produtos, bem como bases auxiliares com informações específicas da Amazon e do Mercado Livre. Além disso, foi implementado um mecanismo de atualização periódica da base principal e do cache de comentários, permitindo que a aplicação opere com dados mais recentes.

No estágio atual, o protótipo já permite operacionalizar funções centrais da proposta do trabalho. Entre elas, destacam-se a filtragem por categoria, a seleção de produtos, a visualização comparativa entre plataformas e a identificação automática da opção de melhor preço. A aplicação também foi estruturada para organizar os comentários associados aos produtos e apresentar, de forma integrada, atributos relevantes para a decisão de compra, como nome do item, preço, nota geral, quantidade de avaliações e link de redirecionamento para a página original do produto.

Além dessas funcionalidades, o protótipo evoluiu para incorporar um módulo de inteligência artificial voltado à síntese automática dos comentários dos consumidores. Essa etapa agregou ao sistema uma análise adicional, na medida em que os comentários coletados passaram a ser utilizados como insumo para a geração de relatórios textuais resumidos. A nova funcionalidade foi implementada com apoio da biblioteca **Hugging Face** e o modelo de IA da Meta, no backend da aplicação, a mesma reforça o caráter aplicado do projeto, ao demonstrar como técnicas contemporâneas de processamento de linguagem natural podem ser integradas a uma plataforma web para ampliar a utilidade prática das informações apresentadas ao usuário.

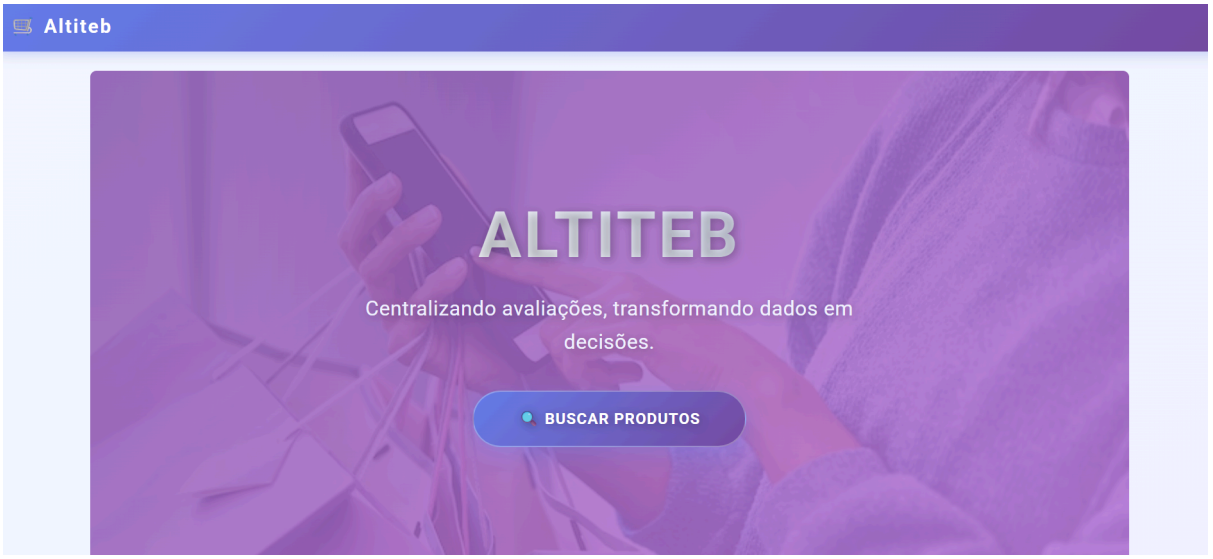
Assim, os resultados alcançados até esta etapa demonstram que o projeto já dispõe de um fluxo funcional completo, iniciando na coleta automatizada, passando pela padronização e integração dos dados, e culminando em uma plataforma web navegável. Esse avanço evidencia a viabilidade técnica da solução proposta e confirma que os componentes

desenvolvidos ao longo da pesquisa conseguem atuar de forma articulada para apoiar a comparação interplataforma de produtos no ambiente do comércio eletrônico.

## 7.2. Interface Atual da Plataforma e Comunicação dos Resultados

O principal resultado tangível do projeto, nesta etapa, é a construção de uma plataforma web funcional voltada à centralização e apresentação comparativa de avaliações de produtos. A interface foi desenvolvida com foco em simplicidade de navegação e clareza visual, de modo a permitir que o usuário percorra, de forma intuitiva, as informações consolidadas a partir das plataformas de e-commerce selecionadas. Diferentemente de uma consulta manual em múltiplos websites, a aplicação reúne em um único ambiente os principais dados necessários para a análise do produto, reduzindo o esforço de busca e comparação.

**Figura 4:** Tela inicial da plataforma.

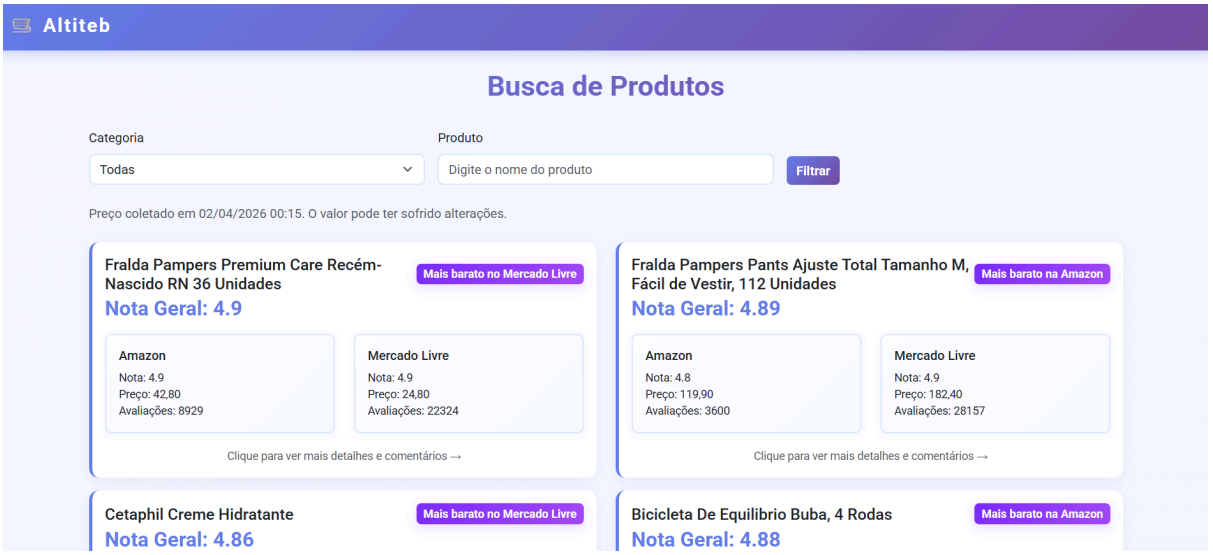



***Fonte:*** *Plataforma Web gerada - Altiteb*

Em sua estrutura atual, a plataforma permite ao usuário iniciar a navegação a partir das categorias, bem como dos produtos referentes a elas, disponíveis na base consolidada. Contando com um total de 124 pares de produtos, distribuídos entre as categorias presentes, a plataforma demonstra uma amostra razoável para a busca dos usuários. A partir da seleção do produto, o sistema exibe uma visão comparativa entre as plataformas, organizando informações como nome do produto, preços encontrados, nota geral, volume de avaliações e comentários associados. Essa funcionalidade analítica transforma dados brutos em

inteligência competitiva para o consumidor, respondendo a perguntas como: "Em qual loja este produto é mais bem avaliado pelos compradores?".

**Figura 5:** Central de busca e listagem comparativa de produtos.




***Fonte:*** *Plataforma Web gerada - Altiteb*

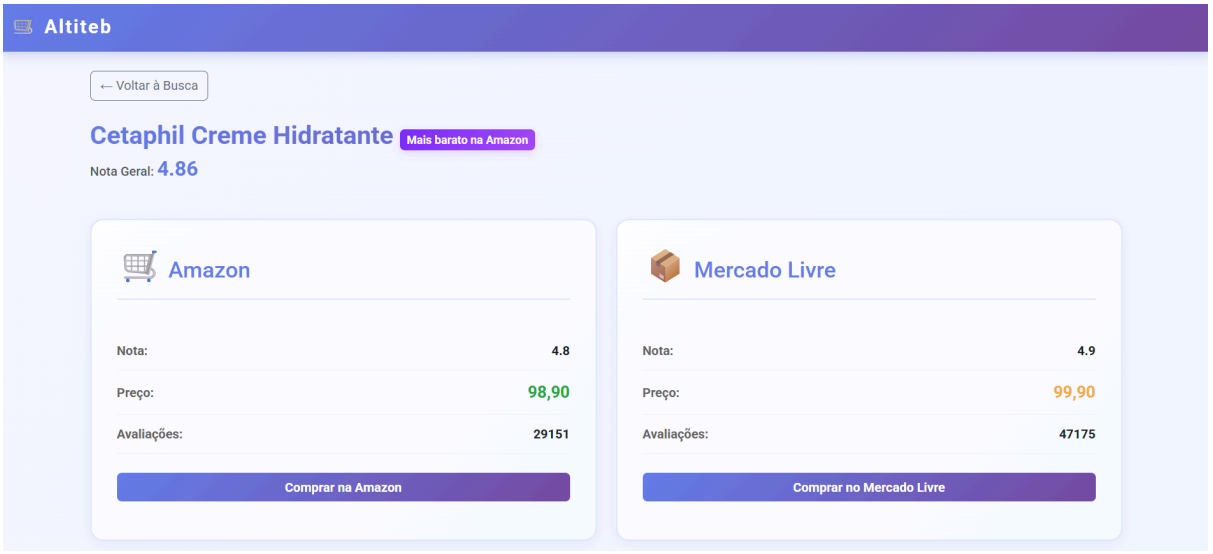
Um aspecto crítico para a credibilidade de qualquer ferramenta que consome dados dinâmicos de e-commerce é a transparência sobre a atualidade das informações apresentadas. Os preços e a disponibilidade de produtos em marketplaces como Amazon e Mercado Livre podem variar diversas vezes ao longo do dia.

Para mitigar esse problema e tornar a ferramenta mais transparente, foi implementado um mecanismo de alerta temporal. Sempre que a aplicação exibe um preço ou uma avaliação, ela informa a data em que aquela informação foi coletada pela última vez, por meio da mensagem: “Preço coletado em [data]. O valor pode ter sofrido alterações.” Essa medida não resolve a defasagem inerente a qualquer sistema baseado em scraping periódico, mas alerta o usuário sobre a necessidade de verificar a informação diretamente na plataforma de origem antes de concluir a compra.

A interface também destaca automaticamente a plataforma que apresenta a melhor condição de preço, funcionando como apoio direto à decisão de compra. Para cada modelo de produto encontrado, como um liquidificador específico, o usuário poderá clicar para acessar um *dashboard* detalhado. Esta tela individualizada é o núcleo da plataforma, onde os dados coletados e processados são apresentados de forma estratégica, tornando assim, a

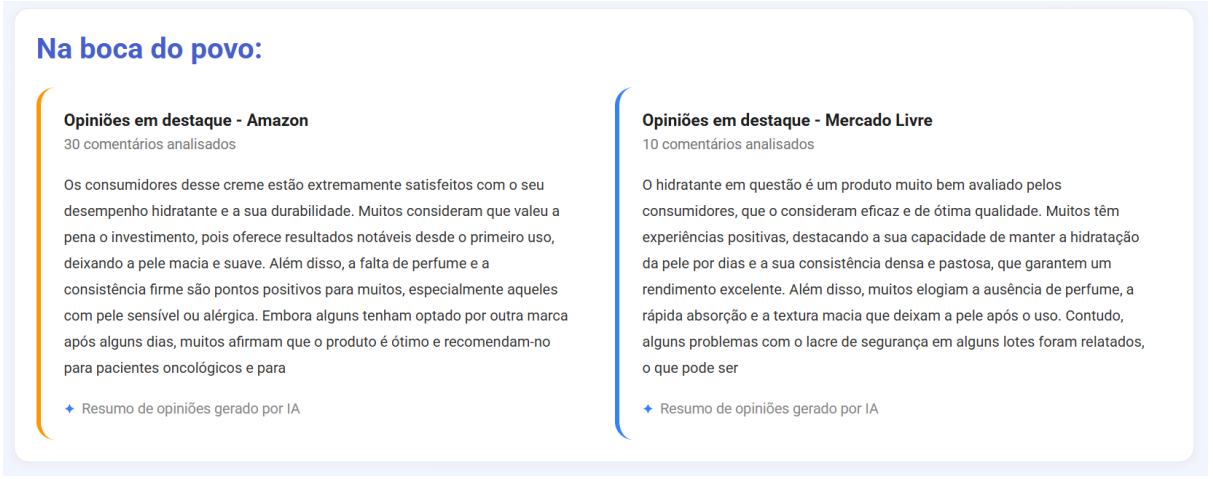
visualização mais objetiva e facilitando a interpretação das diferenças entre as ofertas disponíveis para um mesmo item.

**Figura 6:** Tela de detalhamento comparativo do produto e redirecionamento.




***Fonte:*** *Plataforma Web gerada - Altiteb*

**Figura 7:** Módulo de geração automática de relatório por inteligência artificial.




***Fonte:*** *Plataforma Web gerada - Altiteb*

Como complemento da tela de detalhamento do produto, a aplicação também passou a apresentar um relatório textual gerado por inteligência artificial a partir dos comentários coletados nas plataformas analisadas. Esse recurso foi criado para resumir, em linguagem natural, os principais padrões observados nas avaliações dos consumidores, destacando percepções recorrentes sobre qualidade, desempenho, custo-benefício e eventuais limitações

do produto. Do ponto de vista da experiência do usuário, essa funcionalidade reduz a necessidade de leitura exaustiva de múltiplos comentários individuais, ao transformar um volume de opiniões em uma visão resumida e facilmente interpretável.

**Figura 8:** Tela de comentários e avaliações coletadas.

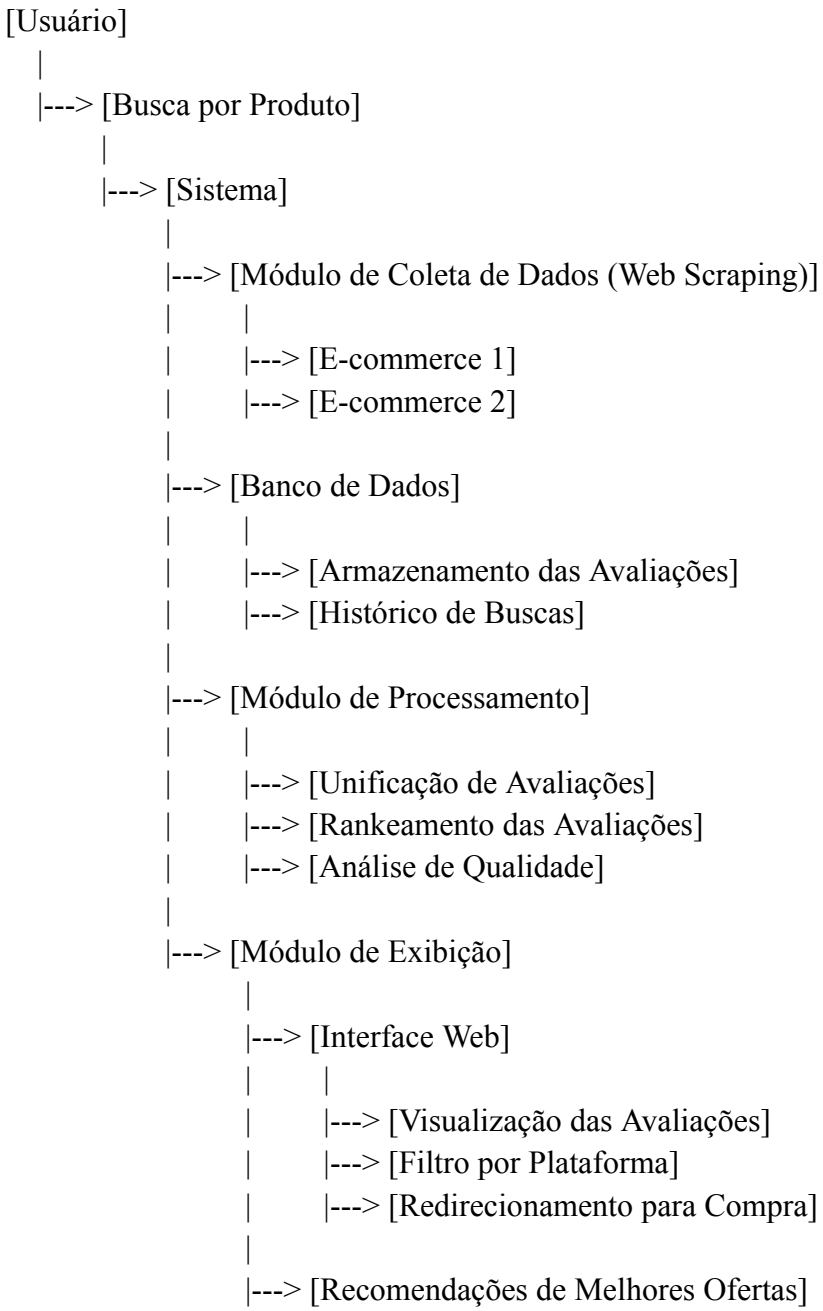

***Fonte:*** *Plataforma Web gerada - Altiteb*

Outro aspecto relevante da aplicação é a integração entre a camada visual e os mecanismos internos de atualização dos dados. O sistema foi preparado para recarregar a base comparativa e atualizar o cache de comentários, o que contribui para manter a consistência das informações apresentadas ao usuário. Além disso, a estrutura do aplicativo foi organizada para consumir diferentes bases derivadas do processo de scraping, incluindo arquivos amostrais, dados completos de comentários e a base master com preços atualizados, consolidando em uma única interface os resultados produzidos nas etapas anteriores do trabalho.

Culminando o processo de análise e decisão, a plataforma oferece uma funcionalidade de redirecionamento, permitindo que, com um clique, o usuário seja levado diretamente à página do produto no site que demonstrou a oferta de maior valor, seja pelo preço, pela robustez das avaliações positivas ou por uma combinação de fatores, fechando o ciclo da jornada de compra de maneira eficiente e informada.

## 7.3. Mapa Conceitual do Sistema

O diagrama a seguir foi projetado para demonstrar os quatro módulos principais (coleta, armazenamento, processamento e exibição), interligados de forma a possibilitar o fluxo completo de informações — desde a busca realizada pelo usuário até a visualização das avaliações unificadas.




Esse modelo conceitual demonstra como os diferentes módulos interagem para entregar ao usuário final uma experiência fluida, centralizada e confiável.

**Descrição dos Componentes:**

- **Usuário:** ator principal, responsável por realizar consultas de produtos na plataforma.

- **Busca por Produto:** mecanismo de entrada no qual o usuário insere palavras-chave do produto desejado.

- **Módulo de Coleta de Dados:** responsável pela extração das avaliações nos sites de e-commerce escolhidos, utilizando scripts de web scraping adaptados a cada plataforma.

- **Banco de Dados:** repositório para armazenamento das avaliações coletadas, metadados dos produtos e cachê de comentários, organizado em formato tabular.

- **Módulo de Processamento:** responsável pela unificação de avaliações de diferentes fontes, pela aplicação de técnicas de ranqueamento e pela análise da qualidade das informações.

- **Módulo de Exibição:** interface web que apresenta os resultados de forma acessível, oferecendo filtros por categoria, comparações entre produtos, redirecionamento para sites de compra e recomendações de melhores ofertas.

# 8. Considerações Finais

O presente trabalho teve como objetivo desenvolver uma solução computacional capaz de coletar, integrar e apresentar de forma unificada avaliações de produtos oriundos de diferentes plataformas de comércio eletrônico, tendo como foco a Amazon e o Mercado Livre. A proposta partiu da constatação de que a fragmentação das avaliações online representa um obstáculo relevante para o consumidor, que frequentemente, precisa consultar múltiplos websites para formar uma percepção mais completa sobre determinado produto. Nesse sentido, o estudo buscou contribuir para a redução dessa dispersão informacional por meio da aplicação de técnicas de web scraping, integração de dados e desenvolvimento web.

Os resultados obtidos demonstram a viabilidade técnica da solução proposta. Foi possível estruturar um pipeline completo de coleta, tratamento, integração e disponibilização dos dados, culminando no desenvolvimento de uma plataforma web funcional. A solução construída permitiu consolidar informações comparativas de 124 produtos mais vendidos da Amazon e do Mercado Livre, reunindo dados como nome, preço, nota geral, quantidade de avaliações e comentários. Além disso, a aplicação desenvolvida possibilita busca por categoria e produto, comparação interplataforma, identificação automática da melhor condição de preço e visualização integrada das informações, tornando mais simples e eficiente o processo de consulta por parte do usuário. Com o avanço adicional do protótipo, destaca-se a incorporação de um módulo de inteligência artificial capaz de gerar relatórios textuais automáticos a partir dos comentários coletados, ampliando o potencial da plataforma ao oferecer uma síntese interpretativa das percepções dos consumidores.

Do ponto de vista metodológico, o trabalho também demonstrou a adequação do uso de técnicas de Processamento de Linguagem Natural para a identificação de produtos equivalentes entre diferentes plataformas. O método baseado em TF-IDF e similaridade de cossenos apresentou desempenho satisfatório para o problema proposto, mostrando-se superior ao método alternativo testado em caráter complementar. Do ponto de vista prático, a pesquisa resultou em um artefato funcional que evidencia a aplicabilidade da ciência de dados e da automação computacional em um problema real do consumo digital.

Como limitações do estudo, destaca-se a delimitação a apenas duas plataformas de comércio eletrônico e a dependência da estrutura dinâmica das páginas para a execução dos scrapers, o que pode exigir ajustes futuros diante de alterações no layout ou nas políticas de

acesso dos sites analisados. Além disso, embora a plataforma já permita a centralização e comparação das avaliações, algumas funcionalidades ainda podem ser aprimoradas, especialmente no que se refere ao enriquecimento analítico dos comentários, e ao tamanho da amostra de produtos.

Uma limitação metodológica adicional, identificada durante a análise dos dados coletados, diz respeito à assimetria nos critérios de exibição de comentários adotados pelas plataformas. A Amazon prioriza as avaliações mais recentes, enquanto o Mercado Livre ordena os comentários por “relevância” , critério que, na prática, privilegia textos mais antigos (por vezes de 2021 ou 2023) em detrimento de opiniões atuais. Esse desalinhamento temporal implica que os resumos gerados por inteligência artificial podem estar sintetizando percepções de épocas diferentes para um mesmo produto, dependendo da plataforma. Consequentemente, a comparação direta de “satisfação atual” entre os dois marketplaces fica comprometida, sendo necessário que o usuário interprete os resultados com ciência dessa heterogeneidade. Futuras versões da ferramenta poderiam contornar essa limitação aplicando filtros temporais uniformes (por exemplo, considerar apenas comentários dos últimos seis meses).

Como perspectivas futuras, o trabalho pode ser expandido pela inclusão de novas plataformas de e-commerce, pelo aprimoramento dos mecanismos de correspondência entre produtos, pela integração com banco de dados e pela evolução da arquitetura da aplicação, com foco em maior robustez, desempenho e escalabilidade. Também se mostra promissora a ampliação do uso de inteligência artificial no sistema, especialmente por meio do aperfeiçoamento dos relatórios automáticos gerados a partir dos comentários, da adoção de técnicas mais avançadas de análise de sentimento e da extração de aspectos específicos das avaliações dos consumidores. Assim, conclui-se que a pesquisa atingiu seu objetivo central e contribuiu tanto academicamente quanto de forma prática para o desenvolvimento de uma solução voltada ao apoio à decisão de compra no comércio eletrônico.

# REFERÊNCIAS


ABCOMM – ASSOCIAÇÃO BRASILEIRA DE COMÉRCIO ELETRÔNICO. **Dados sobre o crescimento do e-commerce no Brasil.** Disponível em: dados.abcomm.org. Acesso em: 27 mar. 2026.

ALBERTIN, A. L. ***Comércio eletrônico: modelo, aspectos e contribuições de sua aplicação.*** 6. ed. São Paulo: Atlas, 2010.

AMAZON WEB SERVICES. **AWS Serverless Multi-Tier Architectures with Amazon API Gateway and AWS Lambda.** [S.l.]: AWS, 2021. Disponível em: documentação oficial da AWS. Acesso em: 27 mar. 2026.

BASS, L.; CLEMENTS, P.; KAZMAN, R. **Software Architecture in Practice.** 3. ed. Boston: Addison-Wesley, 2012.

BHAVSAR, K.; GUPTA, D.; GANATRA, A. **A Comparative Study of Web Scraping Tools.** *International Journal of Computer Applications*, v. 975, n. 8887, p. 1-5, 2018.

BUSCAPÉ. **Cupons de desconto e cashback na Amazon.** Disponível em: https://www.buscape.com.br/cupom-de-desconto/amazon-1582. Acesso em: 28 mar. 2026.

CHEUNG, C. M. K.; THADANI, D. R. **The impact of electronic word-of-mouth communication: A literature analysis and integrative model.** *Decision Support Systems*, v. 54, n. 1, p. 461-470, 2012.

ELMASRI, R.; NAVATHE, S. B. **Fundamentals of Database Systems.** 6. ed. Boston: Pearson, 2016.

FILIERI, R.; MCLEAY, F. **E-WOM and Accommodation: An Analysis of the Factors That Influence Travelers' Adoption of Information from Online Reviews.** *Journal of Travel Research*, v. 53, n. 1, p. 44-57, 2014.

GRONER, R. **HTML, CSS, and JavaScript All in One.** Indianapolis: Sams Publishing, 2018.

HALEVY, A.; RAJARAMAN, A.; ORDONEZ, H. **Data integration: a theoretical perspective.** *VLDB Journal*, v. 18, n. 2, p. 117-125, 2009.

KROTKI, J. **Web Scraping: Legal perspectives and ethical challenges.** *Journal of Information Ethics*, v. 26, n. 1, p. 65-76, 2017.

LAUDON, K. C.; TRAVER, C. G. **E-commerce 2021: business, technology, and society.** 16. ed. Boston: Pearson, 2021.

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

MARTIN, K. **Ethical Issues in Web Data Mining**. In: *Proceedings of the 2015 IEEE International Conference on Big Data*. Washington: IEEE, 2015.

MITCHELL, R. **Web Scraping with Python: Collecting Data from the Modern Web.** 2. ed. Sebastopol: O'Reilly Media, 2018.

RAMALHO, L. **Python Fluente: Programação clara, concisa e eficaz.** 2. ed. Porto Alegre: Bookman, 2022.

RICHARDS, M. **Software Architecture Patterns.** Boston: O'Reilly Media, 2015.

SADALAGE, P. J.; FOWLER, M. **NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence.** Boston: Addison-Wesley, 2012.

SIMILARWEB. **Top websites ranking in Brazil: e-commerce and shopping.** 2026. Disponível em: https://www.similarweb.com/top-websites/brazil/e-commerce-and-shopping/. Acesso em: 28 mar. 2026.

U.S. COMMERCIAL SERVICE. **Brazil - eCommerce.** 2025. Disponível em: https://www.trade.gov/country-commercial-guides/brazil-ecommerce. Acesso em: 28 mar. 2026.

# APÊNDICE A: Quadro de compatibilidade entre categorias da Amazon e do Mercado Livre.

Compatibilidade entre as categorias da Amazon e do Mercado Livre, utilizadas na etapa de similaridade (matching entre produtos).

| Categoria Amazon | Categorias correspondentes no Mercado Livre |
|---|---|
| Alimentos e Bebidas | Alimentos e Bebidas; Esportes e Fitness; Saúde |
| Automotivo | Acessórios para Veículos |
| Bebês | Bebês |
| Beleza | Beleza e Cuidado Pessoal |
| Beleza Premium | Beleza e Cuidado Pessoal |
| Brinquedos e Jogos | Brinquedos e Hobbies |
| Casa | Casa, Móveis e Decoração; Agro |
| CD e Vinil | Música, Filmes e Seriados |
| Computadores e Informática | Informática; Eletrônicos, Áudio e Vídeo; Câmeras e Acessórios |
| Cozinha | Eletrodomésticos; Casa, Móveis e Decoração |
| Dispositivos Amazon e Acessórios | Informática; Eletrônicos, Áudio e Vídeo |
| DVD e Blu-ray | Música, Filmes e Seriados |
| Eletrodomésticos | Eletrodomésticos; Casa, Móveis e Decoração |
| Eletrônicos | Eletrônicos, Áudio e Vídeo; Celulares e Telefones; Câmeras e Acessórios |
| Esporte | Esportes e Fitness; Saúde |
| Ferramentas e Mat. de Construção | Ferramentas; Construção; Agro |
| Games e Consoles | Games |
| Instrumentos Musicais | Instrumentos Musicais |
| Jardim e Piscina | Casa, Móveis e Decoração; Agro |

| | |
|---|---|
| Livros | Livros, Revistas e Comics |
| Moda | Calçados, Roupas e Bolsas; Joias e Relógios |
| Móveis | Casa, Móveis e Decoração |
| Papelaria e Escritório | Arte, Papelaria e Armarinho |
| Pet Shop | Pet Shop |
| Saúde e Bem-Estar | Saúde; Beleza e Cuidado Pessoal; Esportes e Fitness |

***Fonte:*** *Código python para compatibilidade entre categorias*

# APÊNDICE B: Teste complementar de método alternativo de similaridade

Com o objetivo de avaliar a robustez da etapa de correspondência entre produtos das plataformas Amazon e Mercado Livre, foi desenvolvido, em caráter complementar, um método alternativo de similaridade para comparação entre os registros. Diferentemente do método principal adotado no trabalho, essa abordagem baseava-se na extração de atributos diretamente dos nomes dos produtos, como marca, voltagem, cor, capacidade, tamanho e outras características textuais que pudessem contribuir para a identificação de equivalência entre os itens analisados. Após a extração desses atributos, foi definida uma lógica de comparação baseada em pesos, de modo que determinados elementos considerados mais relevantes para a identificação do produto recebessem maior importância no cálculo final de similaridade. A proposta dessa abordagem era tornar o processo de matching mais sensível às características específicas dos produtos, especialmente em situações em que pequenas diferenças textuais pudessem comprometer a comparação direta entre os títulos completos.

Apesar de representar uma tentativa adicional de validação metodológica, os resultados obtidos com esse método não apresentaram desempenho satisfatório em relação ao método principal baseado em TF-IDF e similaridade de cossenos. Na prática, a estratégia por atributos ponderados mostrou menor capacidade de generalização e menor eficácia na identificação de produtos semanticamente equivalentes entre plataformas distintas, resultando em apenas 14 pares de matching, enquanto o método principal adotado no trabalho, retornou um total de 124 pares. A Tabela A1 apresenta uma síntese comparativa entre o método principal de similaridade adotado no trabalho e o método alternativo testado em caráter complementar. O objetivo dessa comparação foi avaliar, de forma resumida, o desempenho

das duas abordagens quanto à identificação de pares equivalentes entre produtos das plataformas Amazon e Mercado Livre.

| **Método** | **Descrição resumida** | **Critério de comparação** | **Pares de matching obtidos** | **Desempenho observado** | **Situação no trabalho** |
|---|---|---|---|---|---|
| TF-IDF + Similaridade de Cossenos | Vetorização dos títulos dos produtos com ponderação pela frequência e raridade dos termos | Similaridade textual entre títulos processados | 124 | Melhor capacidade de generalização e melhor aderência ao problema proposto | Método adotado na solução final |
| Similaridade por Atributos Ponderados | Extração de atributos dos nomes dos produtos e aplicação de pesos diferenciados para comparação | Comparação baseada em atributos textuais relevantes | 14 | Desempenho inferior, com menor eficácia na identificação de produtos equivalentes entre plataformas | Método testado apenas em caráter complementar |

***Fonte:*** *Código python para comparação dos métodos de similaridade*

Dessa forma, o método alternativo não foi incorporado à solução final da plataforma, permanecendo apenas como experimento complementar de verificação. Ainda assim, sua implementação foi importante para reforçar a escolha metodológica adotada no trabalho, uma vez que permitiu comparar abordagens distintas e confirmar que o método principal apresentou melhor aderência ao problema proposto.