UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS / FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Thallys da Silva Nogueira

Observatório do Consumidor: Uma Ferramenta de Mineração de Dados de Redes Sociais para Avaliação de Tendências de Consumo de Derivados Lácteos no Brasil

Juiz de Fora

2025

Thallys da Silva Nogueira

Observatório do Consumidor: Uma Ferramenta de Mineração de Dados de Redes Sociais para Avaliação de Tendências de Consumo de Derivados Lácteos no Brasil

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Modelagem Computacional.

Orientadora: Profa. Dra. Priscila Vanessa Zabala Capriles Goliatt

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

da Silva Nogueira, Thallys.
Observatório do Consumidor: Uma Ferramenta de Mineração de Dados de Redes Sociais para Avaliação de Tendências de Consumo de Derivados Lácteos no Brasil / Thallys da Silva Nogueira. – 2025.
148 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt
Tese de Doutorado – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas / Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional , 2025.

1. Redes Sociais. 2. Pesquisa de Mercado. 3. Análise de Dados. 4. Processamento de Linguagem Natural. 5. Comportamento do Consumidor. I. Capriles, Priscila, orient. II. Título.

Thallys da Silva Nogueira

Observatório do Consumidor: Uma Ferramenta de Mineração de Dados de Redes Sociais para Avaliação de Tendências de Consumo de Derivados Lácteos no Brasil

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 28 de março de 2025.

BANCA EXAMINADORA

Prof.ª Dr.ª. Priscila Vanessa Zabala Capriles Goliatt - Orientadora
Universidade Federal de Juiz de Fora

Prof.ª Dr.ª. Luciana Conceição Dias Campos
Universidade Federal de Juiz de Fora

Prof. Dr. Heder Soares Bernardino
Universidade Federal de Juiz de Fora

Prof.ª Dr.ª. Kennya Beatriz Siqueira
Empresa Brasileira de Pesquisa Agropecuária, Embrapa Gado de Leite

Prof. Dr. Wagner Antônio Arbex

Juiz de Fora, 14/03/2025.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 31/03/2025, às 10:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Heder Soares Bernardino**, **Professor(a)**, em 31/03/2025, às 10:35, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Kennya Beatriz Siqueira**, **Usuário Externo**, em 31/03/2025, às 12:48, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luciana Conceicao Dias Campos**, **Professor(a)**, em 31/03/2025, às 15:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wagner Antonio Arbex**, **Professor(a)**, em 04/04/2025, às 22:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2293184** e o código CRC **8B6D7415**.

Dedico este trabalho a Deus, fonte de força e inspiração, à minha esposa, aos meus pais, aos meus irmãos e aos meus avós. Em especial, dedico à memória de Raulina Maria de Jesus, que, sei, me acompanha do alto em todas as minhas batalhas e conquistas. Meu amor e gratidão eterna a todos vocês!

AGRADECIMENTOS

A Deus, por me conceder força e coragem ao longo de toda essa caminhada. Não foi fácil; houve momentos em que parecia impossível chegar até aqui, mas Sua presença me sustentou. À minha esposa, Luisa Silva Ribeiro Nogueira, pela paciência, compreensão e apoio incondicional ao longo dessa grande jornada. Aos meus pais, Carlos Roberto Nogueira (Beto) e Waldisleia Aparecida da Silva (Kiki), e aos meus irmãos, Thays da Silva Nogueira e Thomás da Silva Nogueira, pelo amor, companheirismo e crença inabalável em mim. Suas palavras de incentivo, encorajamento nos momentos mais difíceis e até as broncas, quando necessárias, foram fundamentais para me moldar e me fortalecer. Vocês foram e sempre serão essenciais para cada conquista em minha vida.

Agradeço, ainda, a todos os meus familiares, que, de alguma forma, estiveram presentes e me apoiaram em cada fase da minha trajetória. Cada gesto de carinho e apoio fez toda a diferença para que eu pudesse alcançar mais esta etapa.

Aos colegas e amigos que tive a felicidade de conhecer ao longo dos cursos de Ciências Exatas, Ciência da Computação, Engenharia Computacional, na Code Empresa Júnior de Computação, no Programa de Pós-Graduação em Modelagem Computacional e em outros cursos da UFJF, meu sincero agradecimento. Também estendo minha gratidão aos colegas e amigos que fiz na Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Gado de Leite.

À Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Gado de Leite de Juiz de Fora, registro meu mais sincero agradecimento por abrir as portas em minha jornada profissional. Foi lá que tive a honra de participar do Programa de Residência Zootécnica Digital, minha primeira grande oportunidade, onde atuei como estagiário e integrante da equipe do "Observatório do Consumidor".

Minha gratidão se estende à pesquisadora, supervisora e uma das idealizadoras do projeto, Kennya Beatriz Siqueira, por sua orientação, confiança e constante incentivo ao longo de todo o projeto.

Um agradecimento especial à minha orientadora, Priscila Vanessa Zabala Capriles Goliatt, cuja orientação, amizade, incentivo e, acima de tudo, paciência e confiança em mim foram fundamentais para a realização deste trabalho. Sem seu apoio e dedicação, nada disso seria possível.

Por fim, expresso minha profunda gratidão à Universidade Federal de Juiz de Fora e à Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Gado de Leite por me proporcionarem a estrutura necessária para meu crescimento pessoal, acadêmico e profissional. Essas instituições foram pilares fundamentais na minha trajetória, oferecendo não apenas oportunidades, mas também momentos inesquecíveis que marcaram minha formação e meu desenvolvimento.

“Comece fazendo o que é necessário, depois o que é possível, e, de repente,
você estará fazendo o impossível!”
São Francisco de Assis.

RESUMO

O avanço da tecnologia tem transformado profundamente a maneira como os dados são coletados, analisados e aplicados em diversas áreas do conhecimento. No campo da pesquisa de mercado, as redes sociais emergem como uma fonte rica e dinâmica de informações, continuamente alimentada por milhões de usuários em todo o mundo. O uso da análise de redes sociais em estudos científicos tem crescido significativamente, impulsionado pelo vasto volume de dados disponíveis e pelo potencial de revelar padrões e tendências de comportamento. Nesse contexto, este estudo propõe o desenvolvimento do Observatório do Consumidor, uma ferramenta inovadora voltada para a pesquisa de mercado no setor lácteo no Brasil, utilizando dados de redes sociais. O objetivo é oferecer uma alternativa eficiente, acessível e representativa aos métodos tradicionais, que, embora cientificamente validados, enfrentam desafios como restrições geográficas, custos elevados e prazos prolongados para aplicação e análise. Para embasar essa proposta, foi conduzida uma revisão sistemática da literatura com o intuito de identificar os principais métodos, análises e técnicas aplicáveis ao desenvolvimento da ferramenta. A partir desses achados e dos princípios de *Business Intelligence*, o Observatório do Consumidor foi criado, explorando dados do X (antigo Twitter), YouTube e Google Trends para extrair informações inéditas sobre o consumo de lácteos no Brasil e suas tendências de mercado. A ferramenta emprega técnicas de processamento de linguagem natural, aprendizado de máquina e banco de dados para realizar análises aprofundadas sobre diferentes aspectos do consumo. Entre suas principais funcionalidades, destacam-se a análise do perfil do consumidor, identificando características como sexo, geração e renda; a análise temporal, que permite compreender a evolução do consumo ao longo do tempo; e a identificação de associações de consumo, revelando quais produtos lácteos e não lácteos são mencionados conjuntamente, evidenciando combinações mais frequentes. Além disso, a ferramenta incorpora uma análise geográfica para mapear a distribuição espacial do consumo e aplica análise comportamental e qualitativa para examinar sentimentos dos consumidores, descrições dos produtos por meio de adjetivos e relatos de experiência, oferecendo uma visão detalhada sobre como o consumo ocorre. Essa abordagem integrada possibilita uma compreensão abrangente do comportamento do consumidor no mercado de lácteos, fornecendo informações estratégicas tanto para a indústria quanto para pesquisadores da área. Os resultados obtidos demonstram o potencial da ferramenta para gerar informações valiosas e inéditas sobre o setor, contribuindo para a tomada de decisões e para o aprimoramento das estratégias de mercado.

Palavras-chave: Redes Sociais. Pesquisa de Mercado. Análise de Dados. Processamento de Linguagem Natural. Comportamento do Consumidor.

ABSTRACT

Technological advancement has profoundly transformed how data is collected, analyzed, and applied across various fields of knowledge. In market research, social media has emerged as a rich and dynamic source of information, continuously fed by millions of users worldwide. The use of social media analysis in scientific studies has grown significantly, driven by the vast volume of available data and its potential to reveal patterns and behavior trends. In this context, this study proposes the development of *Observatório do Consumidor*, an innovative tool for market research in the dairy sector in Brazil, utilizing social media data. The goal is to offer an efficient, accessible, and representative alternative to traditional methods, which, although scientifically validated, face challenges such as geographic restrictions, high costs, and long periods of application and analysis. To support this proposal, we conducted a systematic literature review to identify the main methods, analyses, and techniques applicable to the development of the tool. Based on these findings and the principles of Business Intelligence, *Observatório do Consumidor* was created, exploring data from X (formerly Twitter), YouTube, and Google Trends to extract unprecedented information about dairy consumption in Brazil and its market trends. The tool employs natural language processing techniques, machine learning, and database methods to perform in-depth analyses of various aspects of consumption. Among its key features, the tool includes consumer profile analysis, identifying characteristics such as gender, generation, and income; temporal analysis, which allows an understanding of the evolution of consumption over time; and identification of consumption associations, revealing which dairy and non-dairy products mentions together and highlights the most frequent combinations. Additionally, the tool incorporates geographic analysis to map the spatial distribution of consumption. It applies behavioral and qualitative analysis to examine consumer sentiments and product descriptions through adjectives and experience reports, providing a detailed view of how consumption occurs. This integrated approach enables a comprehensive understanding of consumer behavior in the dairy market, providing strategic insights for industry stakeholders and researchers. The results demonstrate the tool's potential to generate valuable and unprecedented information about the sector, contributing to decision-making and enhancing market strategies.

Keywords: Social Media. Market Research. Data Analysis. Natural Language Processing. Consumer Behavior.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
BERT	*Bidirectional Encoder Representations from Transformers*
BI	*Business Intelligence*
CDC	*Centers for Disease Control and Prevention*
CGU	Conteúdo Gerado pelos Usuários
DP	*Dependency Parsing*
DT	*Decision Tree*
ED	*Event Detection*
EM	*Expectation Maximization*
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IPCA	Índice de Preços ao Consumidor Amplo
IoT	*Internet of Things*
IPEA	Instituto de Pesquisa Econômica Aplicada
ETL	*Extraction, Transformation, Loading*
KNN	*K-Nearest Neighbor*
LDA	*Latent Dirichlet Allocation*
MAE	Erro Absoluto Mediano
MCL	*Markov Clustering*
MAP	Máximo A Posteriori
MSE	Erro Médio Quadrático
MPV	Mínimo Produto Viável
NB	Naive Bayes
NLTK	*Natural Language Processing*
OC	Observatório do Consumidor
ONS	*Office of National Statistics*
PLN	Processamento de Linguagem Natural
POF	Pesquisa de Orçamentos Familiares
POS	*Part-of-speech Tagging*
RBF	*Radial Basis Function*
RMSE	Raiz do Erro Médio Quadrático
SGBD	Sistema de Gerenciamento de Banco de Dados
SEO	*Search Engine Optimization*
SNA	*Social Network Analysis*
SSE	*Sum of Squared Error*
SOC	*Standard Occupational Classification*
SVM	*Support Vector Machine*
TACO	Tabela Brasileira de Composição de Alimentos
TF-IDF	Term Frequency-Inverse Document Frequency
UFJF	Universidade Federal de Juiz de Fora

SUMÁRIO

1 INTRODUÇÃO

1.1 Apresentação do Tema e Contextualização do Problema

Os avanços tecnológicos transformaram profundamente a vida cotidiana, proporcionando melhorias significativas em diversas áreas, especialmente na comunicação. O grande aumento de pessoas conectadas à internet e a popularização das redes sociais resultaram em uma aceleração notável na disseminação de notícias e no compartilhamento de opiniões sobre os mais variados tópicos, incluindo produtos e serviços.

Conforme destacado por CAMBRIA et al. (2013), o processo de extrair e processar de maneira eficiente o vasto volume de informações geradas nesses ambientes tornou-se uma oportunidade estratégica para o mundo empresarial. *Blogs*, redes sociais, sites e fóruns, acessíveis a uma ampla audiência na internet, oferecem percepções valiosas sobre temas como preferências de consumo e tendências de mercado (D'ANDREA et al., 2015). Explorar essas fontes de forma sistemática possibilita decisões mais informadas e direcionadas.

Relatórios publicados pela Datareportal destacam o crescimento constante do número de usuários ativos em redes sociais. Entre outubro de 2018 (DATAREPORTAL, 2018) e outubro de 2022 (DATAREPORTAL, 2022), houve um aumento expressivo de 28,3%, com a base de usuários passando de 3,4 bilhões para 4,7 bilhões. Esses números refletem a crescente adesão de pessoas que interagem diariamente nessas plataformas, evidenciando sua importância como parte integrante da vida moderna.

Com ambientes altamente propícios à interação, as redes sociais podem ser compreendidas como sistemas dinâmicos e complexos. De acordo com SOUZA; QUANDT (2008), esses sistemas são compostos por dois componentes principais: (i) indivíduos que compartilham objetivos e/ou valores semelhantes; e (ii) as conexões que os unem de forma horizontal e descentralizada. Essa estrutura facilita o fluxo de informações e a formação de comunidades interligadas, tornando essas plataformas um campo fértil para estudos e aplicações em diversas áreas.

Entretanto, mesmo com a aparente facilidade de acesso às informações das redes sociais, analisar os dados gerados nesses ambientes apresenta desafios significativos. Essas plataformas produzem dados em formatos variados e frequentemente não estruturados, o que exige técnicas avançadas de processamento para extrair informações relevantes. O Processamento de Linguagem Natural (PLN) desempenha um papel fundamental nesse contexto, permitindo a tradução da linguagem humana, falada ou escrita, em formatos que podem ser utilizados por máquinas (COVINGTON, 1994; RUSSELL; NORVIG, 1995). Essa capacidade é essencial para análise automatizada e eficiente de grandes volumes de texto e interação social.

Uma das características mais marcantes dos dados gerados nas redes sociais é sua natureza não estruturada. Diferentes plataformas apresentam formatos variados de conteúdo, cada uma com peculiaridades que ampliam a diversidade e a complexidade desses dados. O Facebook[1], a maior rede social do mundo, possibilita o compartilhamento de conteúdos em diversos formatos, como imagens, textos, vídeos e *GIFs*, tornando-se uma plataforma versátil para diferentes tipos de interações. Já o YouTube[2], a segunda rede social mais utilizada globalmente, é exclusivamente focado em vídeos, que representam o principal tipo de dado compartilhado na plataforma. Apesar disso, a interação entre os usuários ocorre majoritariamente por meio de comentários e curtidas (*likes*) nos vídeos, configurando uma dinâmica de engajamento distinta. No Instagram[3], o foco recai sobre as imagens, frequentemente acompanhadas de legendas curtas que contextualizam ou complementam o conteúdo visual. Por sua vez, o X[4] (antigo Twitter) se destaca pelo compartilhamento de dados textuais conhecidos como *posts* (que antigamente eram chamados de *tweets*), mensagens curtas limitadas a 280 caracteres, ideais para comunicação rápida e objetiva.

Visando compreender melhor o mercado dos produtos lácteos no Brasil, em 2019 por meio de uma parceria entre a Empresa Brasileira de Pesquisa Agropecuária - Embrapa Gado de Leite, a Universidade Federal de Juiz de Fora e o Instituto Federal do Sudeste Mineiro, surge o Observatório do Consumidor (OC) que é uma ferramenta que inova a maneira de se conduzir pesquisas de mercado tradicionais com o uso de dados de redes sociais para a análise de mercado oferecendo análises e mecanismos capazes de superar as limitações dos métodos tradicionais de pesquisa de mercado (NOGUEIRA, 2021).

Segundo BARABBA; ZALTAMAN (1991), a pesquisa de mercado é definida como o "*processo de ouvir a voz do mercado e transmitir à administração apropriada informações a respeito*", e desempenha um papel crucial no suporte à tomada de decisões. Há décadas, esse método tem sido amplamente utilizado no ambiente empresarial, atingindo seus objetivos e justificando sua aplicação contínua. Essas pesquisas fornecem informações valiosas, como demandas e preferências dos consumidores, estratégias de concorrentes e características de fornecedores, dados essenciais para decisões mercadológicas (BARROS, 2010).

No entanto, em determinados contextos, os métodos tradicionais de pesquisa de mercado podem se revelar ineficientes. Entre os principais desafios estão o elevado número de questões a serem respondidas em questionários, o tempo significativo necessário para aplicação e análise dos resultados além da dificuldade em garantir a representatividade da amostra, sobretudo em pesquisas que exigem a participação de muitos respondentes de diferentes regiões. Esses fatores dificultam a obtenção de dados bem representativos e

[1] https://www.facebook.com. Acesso em: 13 fev. 2025.
[2] https://www.youtube.com. Acesso em: 13 fev. 2025.
[3] https://www.instagram.com. Acesso em: 13 fev. 2025.
[4] https://www.twitter.com. Acesso em: 13 fev. 2025.

distribuídos, comprometendo a eficácia e a aplicabilidade das pesquisas (VERíSSIMO et al., 2018).

Nesse contexto, o OC emerge como uma solução prática e eficiente, capaz de superar tais limitações ao explorar dados de redes sociais de forma ágil, abrangente e representativa. Inicialmente, o OC foi concebido como um Mínimo Produto Viável (MPV), com foco na análise de dados do X relacionados aos queijos artesanais brasileiros. Essa etapa inicial teve como objetivos validar a viabilidade do sistema, identificar informações-chave e compreender os desafios para expandir a análise a outras categorias de produtos lácteos (NOGUEIRA, 2021).

Mas, por que analisar especificamente os produtos lácteos no Brasil, diante da diversidade de temas discutidos nas redes sociais? A resposta está na relevância do leite como uma das *commodities* agropecuárias mais importantes globalmente. Além de seu valor econômico, o leite é uma fonte essencial de nutrientes e amplamente utilizado na culinária e na indústria alimentícia, desempenhando papel central na economia e na cultura alimentar brasileira (SIQUEIRA, 2019).

Adicionalmente, a escassez de informações atualizadas sobre o setor justifica a criação de ferramentas como o Observatório do Consumidor, que permitem analisar o comportamento dos consumidores de forma estratégica. Com isso, é possível subsidiar tanto a indústria quanto os formuladores de políticas públicas na promoção de ações mais eficazes e sustentáveis para o desenvolvimento da cadeia produtiva do leite.

Diante desse contexto, este trabalho teve como objetivo dar continuidade ao desenvolvimento da ferramenta computacional do Observatório do Consumidor[5], promovendo o aprimoramento e a expansão de sua versão inicial. Entre as melhorias implementadas, destacam-se a incorporação de novas fontes de dados, a ampliação do escopo de produtos analisados, a reestruturação do banco de dados e o aperfeiçoamento do modelo de análise de sentimentos. Além disso, foram introduzidas análises mais avançadas e sofisticadas, como a construção de um modelo de análise de sentimentos mais robusto, equilibrado e abrangente, projetado especificamente para os produtos lácteos monitorados pelo observatório.

Outro avanço relevante foi o uso de algoritmos de visão computacional para a extração de características faciais a partir de fotos de perfil, permitindo estimar informações como sexo e faixa etária das pessoas que publicaram sobre produtos lácteos nas redes sociais. Adicionalmente, foram aplicados métodos de análise de dependências sintáticas, que possibilitaram um exame mais detalhado das relações entre palavras nos textos coletados, enriquecendo as análises do consumo de produtos com base nas postagens monitoradas.

Essas melhorias abrangeram todas as etapas do fluxo de análise, incluindo coleta,

[5] *Link* para o sistema do OC: https://observatoriodoconsumidor.cnpgl.embrapa.br. Acesso em: 13 fev. 2025.

armazenamento, processamento e extração de informações dos dados, consolidando o Observatório do Consumidor como uma alternativa eficiente e inovadora às pesquisas de mercado tradicionais (NOGUEIRA, 2021). A ferramenta se mostra eficaz ao possibilitar a compreensão dos anseios e comportamentos dos consumidores por meio da análise de publicações coletadas em redes sociais, contribuindo para a tomada de decisões mais embasadas e alinhadas às demandas do consumidor final.

1.2 Objetivos

Objetivo Geral

Este trabalho tem como objetivo desenvolver e implementar um sistema computacional que integre *Business Intelligence* e análise de dados provenientes de redes sociais.

Objetivos Específicos

- Projetar um processo automatizado de coleta, armazenamento e tratamento de dados oriundos das redes sociais com foco em derivados lácteos no Brasil.
- Modelar um banco de dados relacional escalável capaz de organizar grandes volumes de dados sobre produtos lácteos.
- Construir e validar um *corpus* de treinamento em português do Brasil para o setor de derivados lácteos.
- Aplicar técnicas de mineração textual e linguística computacional para identificar padrões linguísticos relacionados a características, hábitos e preferências de consumo no setor lácteo.
- Avaliar a eficácia do uso de reconhecimento facial e análise de imagens de perfil público como apoio à identificação de perfis de consumidores, respeitando os critérios éticos e legais da pesquisa.
- Desenvolver um sistema *web* com processamento otimizado de consultas a banco de dados e visualização interativa.
- Validar o sistema computacional desenvolvido com base em estudos de caso reais no setor de lácteos no Brasil.

O presente trabalho está organizado da seguinte forma: o capítulo 1 oferece uma breve contextualização do estudo, além de apresentar os objetivos geral e específicos que se propõem a atingir.

O capítulo 2 apresenta uma revisão sistemática da literatura, explorando estudos que utilizam dados de redes sociais como fontes primárias de informação. Esse capítulo aborda os objetivos desses estudos, os métodos aplicados e os algoritmos empregados, destacando tendências e lacunas existentes na literatura.

O capítulo 3 aprofunda os aspectos teóricos dos principais temas abordados nesta pesquisa, incluindo fundamentos relacionados ao uso de dados de redes sociais e técnicas computacionais.

O capítulo 4 detalha os procedimentos metodológicos adotados nesta pesquisa, explicando as etapas necessárias para a implementação dos métodos, a realização das análises e o desenvolvimento do sistema computacional proposto. Essa seção inclui descrições claras e sistemáticas das abordagens utilizadas.

O capítulo 5 é dedicado à apresentação e discussão dos resultados obtidos, analisando os achados à luz da literatura e destacando a contribuição do trabalho para o setor lácteo e para a análise de dados em redes sociais. Além disso, são apresentados trabalhos relevantes que utilizaram o OC como ferramenta de análise.

O capítulo 6 sintetiza as principais contribuições desta pesquisa, destacando os impactos do OC no campo de estudo, evidenciando a relevância e os avanços proporcionados para a área.

O capítulo 7 apresenta as conclusões decorrentes das análises e discussões realizadas ao longo do trabalho. As seções 7.1 e 7.2 abordam, respectivamente, as limitações da ferramenta desenvolvida e sugestões de aprimoramentos futuros. Essas projeções visam expandir as funcionalidades e aplicações do OC, contribuindo para sua evolução contínua e maior relevância no apoio à tomada de decisão.

2 REVISÃO SISTEMÁTICA DA LITERATURA

Neste capítulo, são apresentados em detalhes o procedimento adotado e os resultados obtidos a partir da revisão sistemática da literatura realizada. O principal objetivo foi identificar as técnicas mais relevantes, as ferramentas computacionais e as redes sociais mais utilizadas pela comunidade científica em suas pesquisas, visando identificar elementos que contribuam para o aprimoramento e a compreensão da condução de pesquisas de mercado voltadas à análise do comportamento dos consumidores.

Para estruturar essa revisão, utilizou-se a estratégia PICO[1] (SANTOS et al., 2007), que foi empregada na formulação da questão norteadora da pesquisa, definida como: ***"Quais métodos/ferramentas computacionais podem ser aplicados em dados textuais e de imagem oriundos das redes sociais para identificação de informações do mercado consumidor?"***

Para complementar a construção da resposta a essa questão principal, foram elaboradas seis perguntas secundárias, apresentadas na Tabela 1, com o propósito de detalhar e guiar a coleta e análise dos dados extraídos dos trabalhos incluídos na revisão. Essa abordagem metodológica visou garantir que a análise fosse abrangente e sistemática, possibilitando uma identificação clara das práticas mais relevantes e aplicáveis ao contexto das pesquisas de mercado.

Tabela 1 – Questões secundárias a serem respondidas com a revisão sistemática. **Fonte:** autor.

ID	Pergunta de pesquisa	Objetivo
P1	Quais sãos os contextos de aplicação?	Verificar quais os contextos que cada trabalho possui (e.g., saúde, negócios, política, entre outros)
P2	Quais métodos/ferramentas/análises a comunidade científica aplica em dados de redes sociais?	Descobrir técnicas que podem ser aplicadas a dados de redes sociais para identificação de informações do mercado consumidor(e.g., análise de sentimentos, análise de influência, análise textual, entre outros)
P3	Quais redes sociais são mais utilizadas?	Encontrar as principais plataformas que servem como fonte dados para essas análises (e.g., Facebook, Twitter, Instagram, entre outras)
P4	Quais análises/tratamentos são realizados nos dados?	Identificar os principais procedimentos aplicados aos dados de redes sociais (e.g., limpeza de dados, técnicas de balanceamento, entre outras)
P5	Qual idioma dos dados textuais?	Verificar quais são os idiomas que possuem maior número de aplicação (e.g., inglês, português, entre outros)
P6	Quais são as métricas de avaliação?	Destacar quais métricas de avaliação foram utilizadas para verificar a solução proposta (e.g., *accuracy*, *F1-score*, curva ROC, entre outros)

Para a realização desta revisão sistemática, o Portal de Periódicos CAPES[2] foi utilizado como a principal ferramenta de acesso às bases de dados científicas, por meio

[1] Acrônimo para Paciente, Intervenção, Comparação e "*Outcomes*" (desfecho).
[2] https://www.periodicos.capes.gov.br. Acesso em: 13 fev. 2025.

das credenciais acadêmicas da Universidade Federal de Juiz de Fora. Essa plataforma possibilitou a consulta a uma ampla variedade de artigos científicos, garantindo acesso a fontes confiáveis e de alta relevância para o estudo.

A busca foi conduzida em quatro bases de dados eletrônicas: (1) *ACM Digital Library*[3], (2) *IEEE Digital Library*[4], (3) *Web of Science*[5], e (4) *Scopus*[6]. Essas bases foram selecionadas pela abrangência e qualidade dos conteúdos disponíveis, atendendo aos critérios estabelecidos para a seleção dos artigos.

O procedimento de busca foi estruturado com o seguinte descritor: *((“method” OR “tool”) AND (“data text” OR “data image”) AND “social media” AND “data mining” AND “machine learning”).* Para realizar a coleta de trabalhos na base de dados *Scopus*, foi necessário adicionar um filtro relacionado ao ano de publicação, restringindo a busca a trabalhos publicados entre 2018 e novembro de 2024. Dessa forma, a *string* de busca final utilizada para essa base ficou estruturada da seguinte maneira: *(“method” OR “tool”) AND (“data text” OR “data image”) AND “social media” AND “data mining” AND “machine learning” AND PUBYEAR > 2017).*

Os critérios de inclusão dos trabalhos foram definidos da seguinte maneira: (i) ser categorizado como artigos publicados em *Journals* ou Conferências, dissertações de mestrado ou teses de doutorado; (ii) estudos publicados entre 2018 e novembro de 2024; (iii) possuir alguma das palavras-chave de busca no título ou no resumo; (iv) disponibilidade de acesso ao texto completo; (v) texto escrito em inglês.

Já os critérios de exclusão de trabalhos foram definidos da seguinte forma: (i) idioma do texto do artigo diferente do inglês; (ii) ausência das palavras-chave de busca no título ou no resumo; (iii) trabalhos categorizados como revisões da literatura, livros, capítulos de livros ou títulos de conferências; (iv) apresentar mais de 90% de similaridade com outros trabalhos e (v) não mencionar o uso de dados oriundos de redes sociais.

Para a análise dos trabalhos encontrados, utilizou-se a ferramenta *Rayyan*[7], um aplicativo *web*, gratuito, projetado para auxiliar na confecção de revisões sistemáticas, oferecendo um processo rápido, fácil e conveniente (OUZZANI et al., 2016). Após realizar a busca em cada uma das bases de dados, os arquivos de resultados foram baixados nos formatos *.bibtex*, para *ACM Digital Library*, e *.ris* para as demais bases.

Com os resultados de cada consulta armazenados em seus respectivos arquivos, realizou-se a importação dos dados para o *Rayyan*, iniciando o processo de análise e triagem dos trabalhos. Após submeter todos os arquivos com os resultados de cada uma das bases de dados à plataforma *Rayyan*, obteve-se um total de 461 documentos. Utilizando

[3] https://portal.acm.org/. Acesso em: 13 fev. 2025.
[4] https://ieeexplore.ieee.org/. Acesso em: 13 fev. 2025.
[5] https://www.isiknowledge.com/. Acesso em: 13 fev. 2025.
[6] https://www.scopus.com/. Acesso em: 13 fev. 2025.
[7] https://rayyan.ai/. Acesso em: 13 fev. 2025.

a funcionalidade de detecção de duplicados do *software*, identificaram-se 11 trabalhos duplicados removidos das análises subsequentes. Com isso, 450 trabalhos seguiram para o processo de triagem.

Durante a triagem, 113 trabalhos foram excluídos por não utilizarem dados de redes sociais, 48 por não conterem os descritores de busca no título ou no resumo, e 210 por se tratarem de artigos de revisão da literatura, títulos de conferências, livros ou capítulos de livros. Ao final dessa etapa, restaram 79 trabalhos considerados elegíveis. Dentre esses, 31 foram excluídos devido à indisponibilidade do texto completo, resultando em 48 estudos que atenderam plenamente aos critérios de inclusão e foram incorporados à revisão.

A Figura 1 apresenta o fluxograma do protocolo PRISMA (PAGE et al., 2021), detalhando as etapas realizadas para a seleção dos trabalhos revisados.

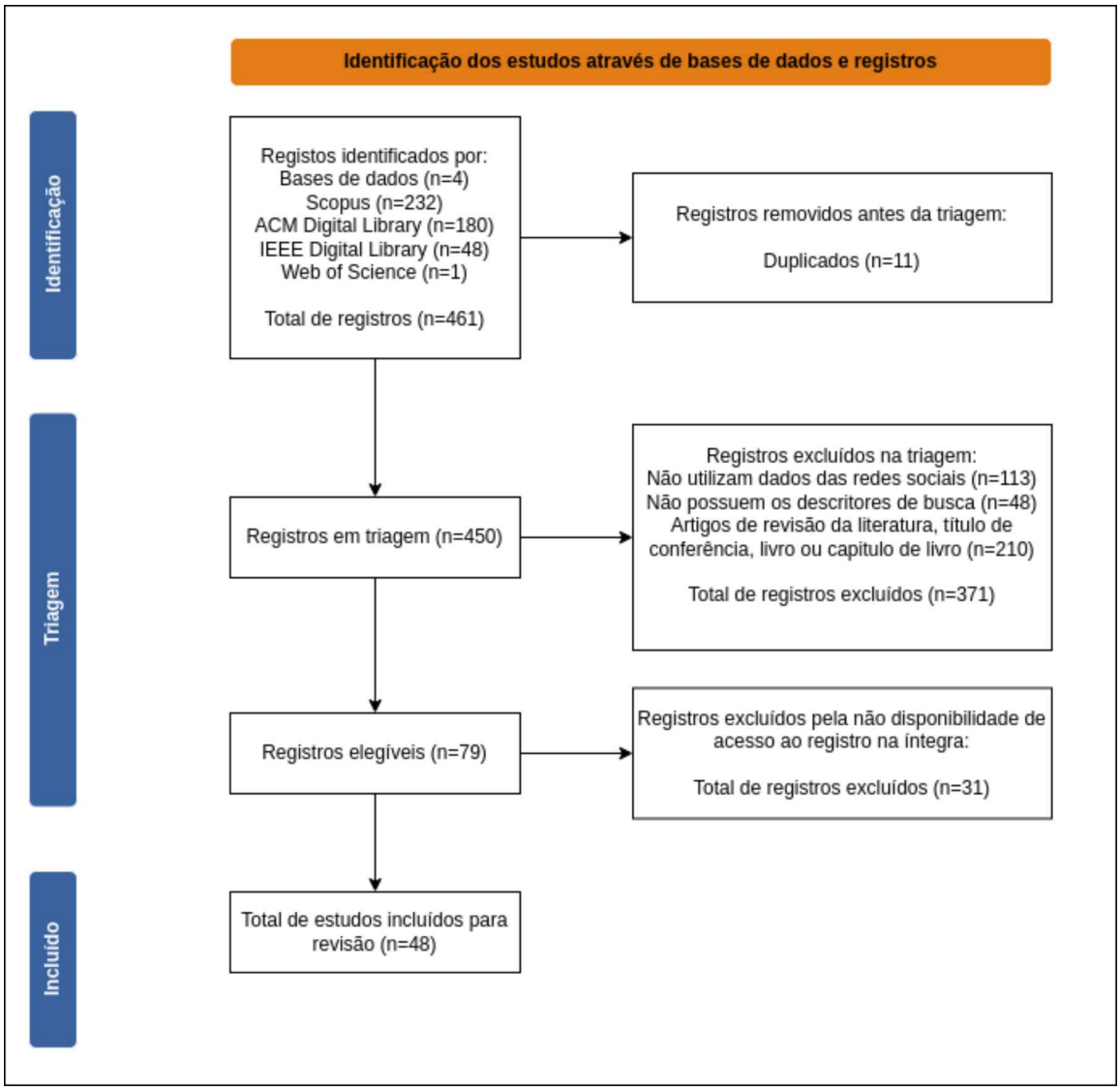


Figura 1 – Fluxograma do protocolo PRISMA utilizado na seleção e triagem dos trabalhos incluídos na revisão sistemática. **Fonte:** autor.

Ao examinar os 48 trabalhos selecionados para leitura integral, foram identificados os dados necessários para, inicialmente, responder às questões secundárias de pesquisa propostas neste estudo. A Tabela 2 exibe as informações dos estudos analisados.

Tabela 2 – Tabela de referências bibliográficas dos estudos incluídos na revisão sistemática. **Fonte:** Autor.

Identificador	**Título do Artigo**	**Referência**
T1	*Enhancing representation in the context of multiple-channel spam filtering*	(NOVO-LOURéS et al., 2022)
T2	*Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry*	(CHANG et al., 2022)
T3	*Identifying comparative opinions in Arabic text in social media using machine learning techniques*	(ALHARBI; KHAN, 2019)
T4	*Does User Generated Content Characterize Millennials' Generation Behavior? Discussing the Relation between SNS and Open Innovation*	(SAURA et al., 2019)
T5	*Improving Arabic sentiment analysis on social media: A comparative study on applying different preprocessing techniques*	(AL-YASIRI; AL-AZAWEI, 2019)
T6	*Analysis of Student Academic Performance and Social Media Activities by Using Data Mining Approach*	(PRATAMA; RIPANTI, 2020)
T7	*Text Categorization of Filipino Tweets Using Naive Byes Algorithm*	(MAGLAPUZ; LACATAN, 2022)
T8	*Social crisis detection using Twitter based text mining-a machine learning approach*	(RAHMAN et al., 2023)
T9	*A Personality Mining System for German Twitter Posts With Global Vectors Word Embedding*	(USSELMANN et al., 2021)
T10	*Change in Threads on Twitter Regarding Influenza, Vaccines, and Vaccination During the COVID-19 Pandemic: Artificial Intelligence-Based Infodemiology Study*	(BENIS et al., 2021)
T11	*A Rumor Detection in Russian Tweets*	(CHERNYAEV et al., 2020)
T12	*Comparing data-driven methods for extracting knowledge from user generated content*	(SAURA et al., 2019b)
T13	*Does SEO Matter for Startups? Identifying Insights from UGC Twitter Communities*	(SAURA et al., 2020)
T14	*Aggregating Twitter text through generalized linear regression models for tweet popularity prediction and automatic topic classification*	(MO et al., 2021)
T15	*How to extract meaningful insights from UGC: A knowledge-based method applied to education*	(SAURA et al., 2019a)
T16	*Identifying startups business opportunities from UGC on Twitter chatting: An exploratory analysis*	(SAURA et al., 2021)
T17	*Implementing Term Frequency-Inverse Term Frequency at Tweets in Indonesian Fraud Crime Cases*	(ULFATRIYANI et al., 2020)
T18	*Making sense of COVID-19 over time in New Zealand: Assessing the public conversation using Twitter*	(JAFARZADEH et al., 2021)

Continua na próxima página

Tabela 2 – Continuação da tabela

Identificador	**Título do Artigo**	**Referência**
T19	*Business Intelligence Model to Analyze Social Media Information*	(KURNIA; SUHARJITO, 2018)
T20	*Making sense of tweets using sentiment analysis on closely related topics*	(BHATNAGAR; CHOUBEY, 2021)
T21	*Marketing challenges in the #MeToo era: gaining business insights using an exploratory sentiment analysis*	(REYES-MENENDEZ et al., 2020)
T22	*A smart Geo-Location Job Recommender System Based on Social Media Posts*	(MUGHAID et al., 2019)
T23	*Multi-Media Content Clustering and Computer Intelligent Analysis by Text Mining*	(REN, 2022)
T24	*Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek*	(JONATHAN et al., 2020)
T25	*Offensive Language Detection in Social Networks for Arabic Language Using Clustering Techniques*	(KANAN et al., 2021)
T26	*An ensemble of deep learning algorithms for popularity prediction of Flickr images*	(ALIJANI et al., 2022)
T27	*Are Black Friday Deals Worth It? Mining Twitter Users' Sentiment and Behavior Response*	(SAURA et al., 2019c)
T28	*Sentiment analysis for Arabic social media news polarity*	(HNAIF et al., 2021)
T29	*Social media toxicity classification using deep learning: Real-world application UK Brexit*	(FAN et al., 2021)
T30	*The Effect of Platform Intervention Policies on Fake News Dissemination and Survival: An Empirical Examination*	(NG et al., 2022)
T31	*Tweet normalization: A knowledge-based approach*	(GUPTA; JOSHI, 2017)
T32	*Understanding the public discussion about the centers for disease control and prevention during the COVID-19 pandemic using Twitter data: Text mining analysis study*	(LYU; LULI, 2021)
T33	*Using data mining techniques to explore security issues in smart living environments in Twitter*	(SAURA et al., 2021)
T34	*Using social media audience data to analyze the drivers of low-carbon diets*	(EKER et al., 2021)
T35	*Novelty Detection in Social Media by Fusing Text and Image into a Single Structure*	(AMORIM et al., 2019)
T36	*Adaptive Binary Bat and Markov Clustering Algorithms for Optimal Text Feature Selection in News Events Detection Model*	(AL-DYANI et al., 2022)
T37	*Utilization of social media in floods assessment using data mining techniques*	(KHAN et al., 2022)
T38	*Classification of Tourism Reviews from Bengali Texts using Multinomial Naïve Bayes*	(BAPPON; IQBAL, 2022)
T39	*An Exploratory Study of Tweets about the SARS-CoV-2 Omicron Variant: Insights from Sentiment Analysis, Language Interpretation, Source Tracking, Type Classification, and Embedded URL Detection*	(THAKUR; HAN, 2022)

Continua na próxima página

Tabela 2 – Continuação da tabela

Identificador	**Título do Artigo**	**Referência**
T40	*Flood Depth Assessment with Location-Based Social Network Data and Google Street View: A Case Study with Buildings as Reference Objects*	(ZOU et al., 2022)
T41	*Investigating and Analyzing Self-Reporting of Long COVID on Twitter: Findings from Sentiment Analysis*	(THAKUR, 2023)
T42	*Development of Context-Based Sentiment Classification for Intelligent Stock Market Prediction*	(SMATOV et al., 2024)
T43	*A corpus-based real-time text classification and tagging approach for social data*	(MEMON et al., 2024)
T44	*Impact of Preprocessing on Twitter Based Covid-19 Vaccination Text Data by Classification Techniques*	(SRIDEVI; VELMURUGAN, 2022)
T45	*What Remains Now That the Fear Has Passed? Developmental Trajectory Analysis of COVID-19 Pandemic for Co-occurrences of Twitter, Google Trends, and Public Health Data*	(RATHKE et al., 2023)
T46	*Hoax Information Detection System Using Apriori Algorithm and Random Forest Algorithm in Twitter*	(UTAMI et al., 2020)
T47	*Twitter Sentiment Analysis of the Low-Cost Airline Services After COVID-19 Outbreak: The Case of AirAsia*	(SAAD et al., 2023)
T48	*Construction of a training dataset for a sentiment analysis model of dairy products tweets in Brazil*	(NOGUEIRA et al., 2024)

Fim da tabela

2.1 Respostas das Questões Secundárias de Pesquisa

Respondendo à primeira questão secundária de pesquisa **P1: Quais são os contextos de aplicação?** Foi possível ter a percepção de que os contextos dos trabalhos foram os mais variados possíveis. Em cada trabalho analisado verificou-se que as redes sociais tiveram papel fundamental na condução das pesquisas. Para proporcionar uma visão geral dos contextos abordados, os estudos foram organizados em 7 eixos temáticos com base nos tipos de assunto tratados que são: (i) Tecnologia, métodos e modelos; (ii) Informação, comunicação e redes sociais; (iii) Saúde pública; (iv) *Marketing*, mercado e negócios; (v) Turismo, (vi) Educação e (vii) Monitoramento de desastres ambientais. A Figura 2 exibe em detalhes os trabalhos que pertencem a cada um dos eixos temáticos citados.

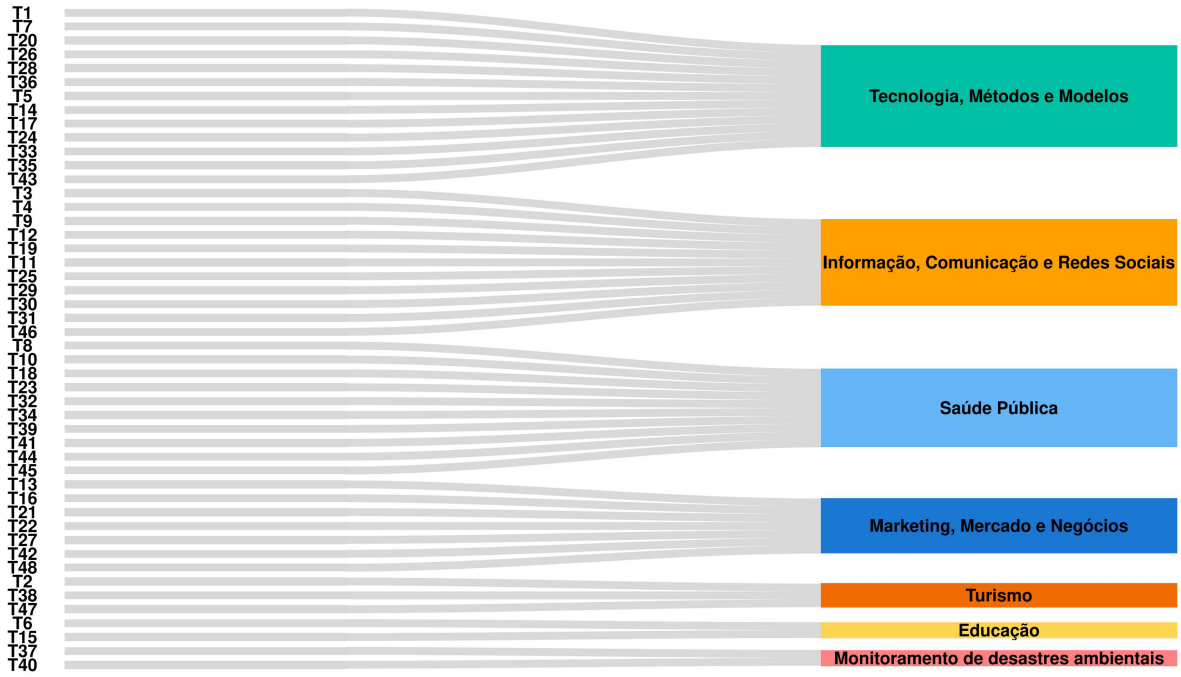


Figura 2 – Esquema ilustrativo dos eixos temáticos e contextos abordados nos trabalhos analisados na revisão sistemática. **Fonte:** autor.

2.1.1 Análises de cada eixo temático

2.1.1.1 Tecnologia, métodos e modelos

É possível observar no resultado apresentado que a distribuição dos eixos temáticos reflete a diversidade de abordagens nos trabalhos analisados, com destaque para os eixos relacionados a áreas mais amplas e contemporâneas, o eixo de Tecnologia, Métodos e Modelos (13) foi o que obteve maior número de trabalhos categorizados nesta classe. Os estudos abrangem subtemas variados, sendo a modelagem e o desenvolvimento de algoritmos avançados os mais recorrentes. Esses trabalhos destacam o uso de técnicas sofisticadas para solucionar problemas complexos em diferentes áreas, como filtragem de *spam*, análise de sentimentos e identificação de padrões em grandes volumes de dados.

NOVO-LOURéS et al. (2022) e HNAIF et al. (2021) focaram no processamento de textos para o desenvolvimento de tarefas específicas. O primeiro propôs uma ferramenta para identificação de mensagens *spam*, enquanto o segundo construiu um *corpus*[8] manualmente anotado em árabe para análise de sentimentos.

Outros trabalhos voltaram-se à análise de sentimentos e compreensão de interações sociais. BHATNAGAR; CHOUBEY (2021) desenvolveram uma arquitetura que combina análise de sentimentos e detecção de comunidades, permitindo identificar sentimentos gerais em tópicos inter-relacionados, enquanto SAURA et al. (2021) aplicaram análise de sentimentos e modelagem de tópicos para explorar questões de segurança em ambientes de

[8] Um *corpus* é uma coletânea de textos naturais, escolhidos para caracterizar um estado ou variedade de linguagem. (SARDINHA, 2000)

vida inteligente.

Focando em redes sociais e dados multimodais[9], ALIJANI et al. (2022) introduziram um modelo híbrido para prever a popularidade de conteúdos, integrando dados de imagem e temporais normalizados, enquanto AL-DYANI et al. (2022) aplicaram o *Binary Bat Algorithm* adaptado, combinado com *Markov Clustering* (MCL), para selecionar características relevantes na detecção de eventos em textos heterogêneos.

Ainda dentro deste eixo, foram observados estudos relacionados com tarefas de predição e classificação para categorizar e etiquetar dados sociais. MAGLAPUZ; LACATAN (2022) apresentaram um modelo baseado em *Naive Bayes* para classificar *tweets* filipinos em categorias como política, tecnologia, entretenimento e esportes. Já MO et al. (2021) propuseram um método para classificar *tweets* e avaliar sua popularidade. Por sua vez, MEMON et al. (2024) desenvolveram uma abordagem para classificação e etiquetagem de dados sociais em tempo real, focada em semelhanças semânticas e morfológicas. Outro problema abordado foi o de explorar técnicas para tratar dados desbalanceados (JONATHAN et al., 2020), que é um problema muito frequente em dados nesses tipos de tarefas.

Os trabalhos de AL-YASIRI; AL-AZAWEI (2019), ULFATRIYANI et al. (2020) e AMORIM et al. (2019) abordaram diferentes aspectos da análise de dados nas redes sociais. AL-YASIRI; AL-AZAWEI (2019) investigaram o impacto de diferentes formas de processamento dos dados para modelos de análise de sentimentos em árabe. ULFATRIYANI et al. (2020) focaram na identificação de fraudes usando *tweets* em indonésio usando a técnica TF-IDF para detectar padrões, embora enfrentaram limitações com a escassez de dados no idioma. AMORIM et al. (2019) propuseram uma abordagem para detectar tópicos através da fusão de dados heterogêneos (imagem + texto) utilizando redes neurais convolucionais.

2.1.1.2 Informação, comunicação e redes sociais

O eixo Informação, Comunicação e Redes Sociais (11), aparece como o segundo eixo mais representativo, indicando um forte interesse em compreender os impactos e as dinâmicas das redes sociais nos contextos de cada um dos estudos.

Neste eixo pode-se destacar trabalhos focados em detectar e tratar padrões textuais (GUPTA; JOSHI, 2017) como gírias, *hashtags* e *emoticons* muito presentes em textos de redes sociais, avaliação de aspectos comportamentais de usuários da geração *Millennials*[10] (SAURA et al., 2019) e até o desenvolvimento de modelos capazes de identificar traços de

[9] Um modelo multimodal é um sistema de inteligência artificial projetado para integrar e interpretar informações provenientes de diferentes modalidades de dados, como texto, imagens, áudio, vídeo ou dados de sensores (ULTRALYTICS, 2025).

[10] Os *Millennials*, ou Geração Y, são as pessoas nascidas entre 1981 e 1995, predominam no mercado de trabalho, destacando-se como nativos digitais, inovadores e idealistas (URCO et al., 2019).

personalidade de um usuário (USSELMANN et al., 2021) por meio da análise do conteúdo gerado por eles nas redes sociais.

Outra vertente de estudos deste eixo focou suas análises na tentativa de detectar as chamadas *fake news*, ou notícias falsas, amplamente disseminadas em mídias sociais, além de conteúdos ofensivos. Assim, as pesquisas buscaram identificar padrões nos dados da rede social X, para detectar informações falsas disseminadas (UTAMI et al., 2020; CHERNYAEV et al., 2020). Complementando essa abordagem, NG et al. (2022) avaliaram o impacto da aplicação de políticas de intervenção das plataformas na redução da disseminação de notícias falsas. Além disso, KANAN et al. (2021) direcionaram seus estudos para a detecção de linguagem ofensiva em redes sociais, com foco específico em dados na língua árabe enquanto FAN et al. (2021) abordaram a toxicidade em redes sociais, utilizando modelos baseados em BERT[11] e alcançando alta precisão ao classificar comentários tóxicos em casos como o *Brexit*[12] no Reino Unido.

Outros trabalhos dentro desse eixo focaram na extração de conhecimento empresarial por meio da aplicação de técnicas de mineração de dados integradas a conceitos de *Business Intelligence* (BI). Esses estudos demonstram como essas abordagens podem oferecer soluções práticas e eficazes, abrangendo desde a análise de opiniões (ALHARBI; KHAN, 2019) até a identificação de tópicos e tendências em redes sociais (SAURA et al., 2019b; KURNIA; SUHARJITO, 2018). Esses resultados destacam o potencial dessas técnicas em apoiar a tomada de decisões estratégicas, muito importantes para o planejamento e gestão empresarial.

2.1.1.3 Saúde pública

A análise dos trabalhos relacionados ao eixo temático de Saúde Pública (10) revelou um esforço significativo durante a pandemia de COVID-19 para compreender, por meio das redes sociais, as mudanças na percepção pública sobre a doença, utilizando essas plataformas como fonte de dados para a análise de crises e comportamentos sociais. Conforme destacado por NOGUEIRA et al. (2024a), a pandemia impactou profundamente a rotina das pessoas, transformando as mídias sociais em importantes válvulas de escape diante do desgaste mental agravado pelas notícias sobre mortes e pela intensa busca pela vacina.

Nesse contexto, o estudo de RAHMAN et al. (2023) investigou crises sociais causadas pela pandemia em Bangladesh, usando dados do Twitter comparados com jornais

[11] O *Bidirectional Encoder Representations from Transformers* (BERT) é um modelo de linguagem desenvolvido por Jacob Devlin e sua equipe na *Google AI Language* em 2018 (DEVLIN et al., 2019). Ele revolucionou o pré-treinamento de representações em linguagem natural ao introduzir uma abordagem bidirecional, permitindo a captura mais precisa dos contextos linguísticos.

[12] O *Brexit* é um termo utilizado para descrever a saída do Reino Unido da União Europeia.

confiáveis, identificando crises via análises de sentimentos negativos. Similarmente, BENIS et al. (2021) examinaram discussões sobre vacinação contra *influenza* e COVID-19 no Twitter, destacando como o monitoramento de redes sociais pode influenciar campanhas de vacinação. Na Nova Zelândia, JAFARZADEH et al. (2021) analisaram conversas no Twitter para entender as reações da população a medidas proativas de controle da pandemia, enquanto REN (2022) utilizou dados da plataforma Sina Weibo para estudar reações e disseminação de informações durante a pandemia, ajudando a ajustar respostas governamentais. Já LYU; LULI (2021) exploraram as discussões públicas sobre o CDC, sigla para *Centers for Disease Control and Prevention*, no Twitter, identificando preocupações e percepções sobre diretrizes e credibilidade institucional estratégicas para a comunicação de saúde pública.

O trabalho de THAKUR; HAN (2022) focou em dados relacionados à variante *Omicron*, analisando os sentimentos e linguagem expressos nos *tweets* para entender engajamentos com informações sobre a pandemia. Da mesma forma, THAKUR (2023) utilizaram o Twitter para estudar experiências relacionadas à COVID longa, destacando a relevância de captar percepções públicas para formulação de políticas de saúde. RATHKE et al. (2023) investigaram como as emoções e percepções públicas mudaram ao longo da pandemia, combinando dados do Twitter, Google Trends e informações de saúde pública para entender dinâmicas de opinião, polarização política e disseminação de desinformação. Por fim, SRIDEVI; VELMURUGAN (2022) investigaram a eficácia de técnicas de pré-processamento e algoritmos de classificação para analisar sentimentos sobre a vacinação contra a COVID-19 no Twitter.

Com outro foco, o estudo de EKER et al. (2021) mostrou a utilização de dados de segmentação de audiência da plataforma publicitária do Facebook para analisar a extensão e os motivadores de interesse em estilos de vida sustentáveis, especialmente de dietas baseadas em plantas, ao nível global. Esses estudos evidenciam a crescente aplicação de tecnologias emergentes em saúde pública, enfatizando sua relevância em um cenário global onde inovação e desafios de saúde estão cada vez mais interconectados.

2.1.1.4 *Marketing*, mercado e negócios

No eixo de *Marketing*, Mercado e Negócios (7), os estudos destacam o potencial das redes sociais para fornecer informações sobre comportamento do consumidor, tendências de mercado e estratégias empresariais. SAURA et al. (2020) exploraram a relevância de *Search Engine Optimization* (SEO) para *startups*, identificando comunidades no Twitter que discutem otimização de mecanismos de busca e extraindo indicadores-chave baseados nos sentimentos expressos nas publicações. Já SAURA et al. (2021) analisaram o ecossistema de *startups* indianas, identificando oportunidades de investimento em setores como *fintech*, inovação e inteligência artificial, utilizando modelagem de tópicos e análise de sentimentos

para revelar o impacto desses fatores no crescimento exponencial do mercado.

No contexto dos movimentos sociais, REYES-MENENDEZ et al. (2020) investigaram o impacto do movimento #MeToo no *marketing* e nas campanhas publicitárias, evidenciando a relevância da igualdade de gênero e da inclusão nas comunicações corporativas. Complementarmente, SAURA et al. (2019c) analisaram o comportamento dos consumidores espanhóis durante a *Black Friday* de 2018, ressaltando percepções positivas em relação a promoções exclusivas e produtos tecnológicos, enquanto identificaram reações negativas associadas ao suporte ao cliente e a práticas fraudulentas.

Outro enfoque inovador foi apresentado por MUGHAID et al. (2019), que desenvolveram um sistema de recomendação de empregos baseado em geo-localização, utilizando informações de redes sociais para conectar candidatos a oportunidades ideais. Já SMATOV et al. (2024) propuseram um modelo de análise de sentimentos voltado para prever movimentos do mercado de ações, utilizando redes neurais para capturar relações semânticas diretamente dos dados textuais. Por fim, NOGUEIRA et al. (2024) desenvolveram um modelo de análise de sentimentos voltado especificamente para o setor lácteo no Brasil. O estudo envolveu a coleta de *tweets* para a construção de um modelo balanceado, capaz de classificar dados inéditos. Além disso, essa aplicação foi integrada a um sistema computacional que fornece análises em tempo real sobre o comportamento do consumidor e as tendências de mercado.

2.1.1.5 Turismo

No eixo de Turismo (3), os estudos evidenciam o potencial da análise de sentimentos como uma ferramenta estratégica para compreender a satisfação dos clientes e aprimorar a experiência no setor. CHANG et al. (2022) analisaram avaliações de voos na TripAdvisor entre 2016 e 2020, investigando o impacto da COVID-19 na percepção dos passageiros. Os resultados indicam que a satisfação ou insatisfação dos viajantes está diretamente ligada à comparação entre expectativas e experiências, com variações significativas conforme a classe da cabine e o preço dos bilhetes. Por sua vez, BAPPON; IQBAL (2022) desenvolveram um sistema automatizado baseado em aprendizado de máquina para classificar avaliações turísticas no idioma bengali em categorias positivas e negativas, auxiliando turistas na escolha de destinos com base em análises de sentimentos.

No estudo de SAAD et al. (2023) analisaram *tweets* sobre os serviços da AirAsia, identificando uma predominância de sentimentos negativos entre os passageiros. O estudo destacou a necessidade de melhorias nos serviços, integrando modelos de satisfação do cliente e dimensões de qualidade digital para compreender melhor as emoções e expectativas dos consumidores.

2.1.1.6 Educação

No eixo da educação (2), dois trabalhos se destacaram ao explorar a interseção entre tecnologia, comportamento social e desempenho acadêmico. O primeiro, conduzido por PRATAMA; RIPANTI (2020), analisou a relação entre dados acadêmicos e interações em redes sociais. Foram coletados dados de 20 estudantes, incluindo número de *posts*, quantidade de seguidores, conteúdo textual das publicações e datas de criação das publicações. Os resultados revelaram uma correlação interessante: estudantes com notas médias abaixo de 3,00 tendiam a ter mais seguidores em redes sociais do que aqueles com notas mais altas, sugerindo que o engajamento digital pode estar relacionado a padrões de desempenho acadêmico.

Por sua vez, o estudo de SAURA et al. (2019a) enfatiza o papel das plataformas digitais na compreensão das interações e experiências de alunos e professores. Utilizando técnicas de análise de dados, os autores extraíram informações significativas voltadas para melhorar a comunicação, o *marketing* educacional e, sobretudo, a experiência de aprendizagem. O trabalho destaca como dados digitais podem ser utilizados para promover ambientes educacionais mais eficazes e alinhados às necessidades de seus usuários.

2.1.1.7 Monitoramento de desastres ambientais

Por fim, no eixo de Monitoramento de desastres ambientais (2), os estudos exploram o uso de tecnologias e dados de mídias sociais para analisar eventos de inundações. KHAN et al. (2022) sugerem um método alternativo para monitorar inundações utilizando dados multimodais (texto, imagens e vídeos) coletados de plataformas como Facebook, Twitter e YouTube. O estudo focou em eventos nos Emirados Árabes Unidos, avaliando a confiabilidade e qualidade dos dados em regiões áridas com acesso limitado a medidores de vazão. O trabalho demonstrou uma correlação significativa entre dados de mídias sociais e registros pluviométricos, validando a utilidade dessas fontes alternativas.

Já ZOU et al. (2022) propuseram um método para avaliar desastres de inundações em tempo real, combinando imagens de redes sociais com dados do *Google Street View*. Por meio da segmentação de edifícios e análise visual, os autores conseguiram estimar a profundidade da água, superando as limitações de métodos tradicionais que dependiam de dados caros e escassos. A abordagem foi testada em desastres no Japão e nos Estados Unidos, demonstrando sua eficácia e viabilidade prática.

Em conclusão, a análise da distribuição dos eixos temáticos revelou que os dados de redes sociais são amplamente aplicados em diversos contextos, com uma forte ênfase nas áreas de tecnologia e sua integração com questões sociais. Diversos estudos realizaram comparações entre os resultados obtidos a partir de dados de redes sociais e informações de outros meios de comunicação, identificando correlações significativas entre essas bases de dados. Contudo, observou-se que o tema relacionado às características, hábitos e consumo

de alimentos, particularmente os derivados lácteos, é um campo bastante específico e pouco explorado, sendo abordado apenas no trabalho de NOGUEIRA et al. (2024) e ausente nos demais estudos desta revisão. Esse fato destaca a relevância e o potencial de desenvolver uma ferramenta dedicada à análise de dados nesse contexto, possibilitando a obtenção de informações inéditas sobre o mercado de lácteos no Brasil.

Dando continuidade às análises, ao recuperar e examinar os dados necessários para responder à segunda questão, **P2. Quais métodos, ferramentas e análises a comunidade científica aplica em dados de redes sociais?**, identificou-se uma ampla diversidade de abordagens utilizadas para explorar esses ambientes virtuais. Entre elas, a análise de sentimentos destacou-se como a mais empregada, sendo adotada em 27,16% dos trabalhos analisados. Em seguida, a análise textual apareceu em 18,52% dos estudos, evidenciando sua relevância na identificação de padrões linguísticos e semânticos em dados textuais.

Outro destaque foi a modelagem de tópicos, empregada em 16,05% dos estudos, evidenciando o interesse em identificar temas latentes nas discussões presentes nas redes sociais. Métodos voltados à previsão foram aplicados em 4,94% dos trabalhos. De forma similar, a análise estatística também apareceu em 4,94% dos estudos. As técnicas de classificação, utilizadas para categorizar informações em diferentes classes ou grupos, foram aplicadas em 6,17% dos trabalhos. Já métodos de clusterização e detecção de comunidades, usados para agrupar dados ou identificar estruturas em redes sociais, apresentaram representatividade de 3,70% cada.

Abordagens como análise geográfica, análise temporal, categorização de textos, detecção de características, mineração de dados, processamento de textos, redes complexas, relacionamento de dados, sistemas de recomendação, *web scraping*, visão computacional e modelagem matemática, tiveram aplicação mais restrita, cada uma representando 1,23% dos trabalhos analisados. Esses dados revelam a diversidade de enfoques utilizados pela comunidade científica para extrair informações valiosas de redes sociais, mostrando como os métodos e ferramentas variam conforme o objetivo da pesquisa e a natureza dos dados analisados.

Com relação à questão **P3. Quais redes sociais são mais utilizadas?**, a Figura 3 mostra os resultados obtidos.

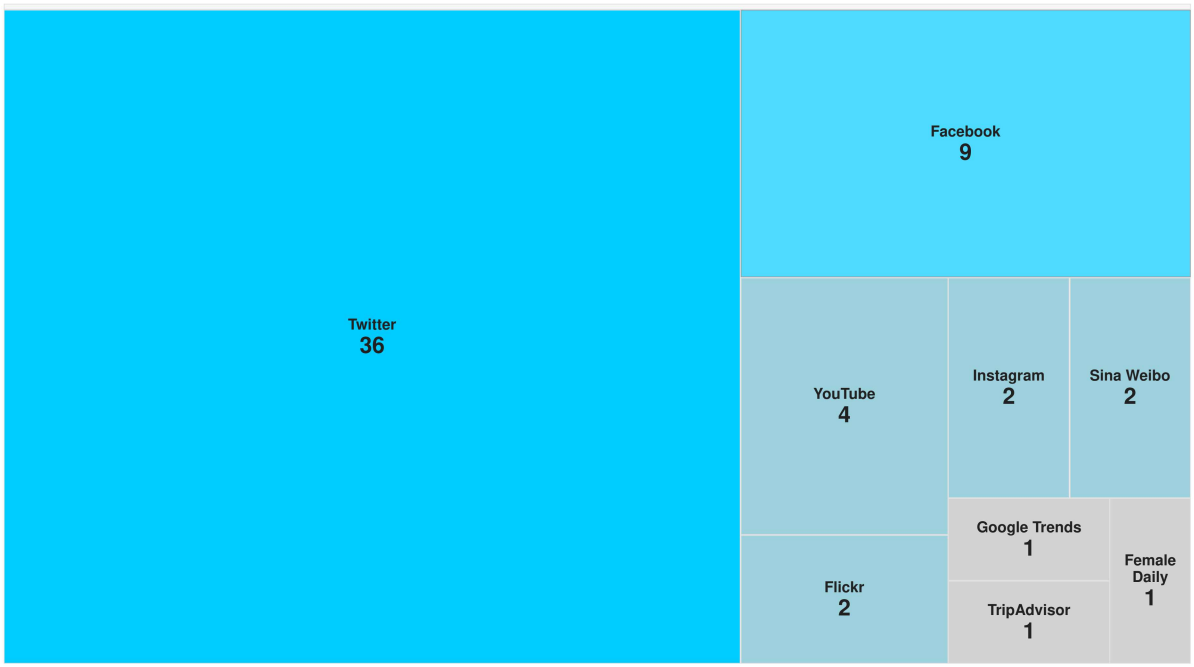


Figura 3 – Esquema ilustrativo das redes sociais mais usadas nos trabalhos analisados na revisão sistemática. **Fonte:** autor.

É possível verificar que o Twitter foi a rede social mais empregada pela comunidade científica, presente em 62,07% dos trabalhos analisados. A segunda rede mais utilizada foi o Facebook, com 15,52%, seguida pelo YouTube, com 6,9%. Já as redes Flickr, Instagram e Sina Weibo foram mencionadas em 3,45% dos estudos, enquanto TripAdvisor, Female Daily e Google Trends apareceram em 1,72% dos trabalhos.

Esses resultados são particularmente relevantes, pois, de acordo com o relatório publicado por DATAREPORTAL (2022), em janeiro de 2022, o Twitter ocupava a 15ª posição entre as redes sociais mais populares globalmente, com 436 milhões de usuários ativos. Em outubro de 2024, o relatório da DATAREPORTAL (2024) apontou que, desde a sua mudança de gestão e de nome para X em 2023, a plataforma não divulga mais o número de usuários ativos, informando apenas seu alcance potencial de anúncios, estimado em 590 milhões de usuários. Embora expressivo, esse número permanece consideravelmente inferior ao das principais redes sociais, como Facebook (3,07 bilhões de usuários ativos mensais), YouTube (com alcance potencial de 2,53 bilhões), WhatsApp e Instagram (2 bilhões de usuários ativos cada), TikTok (com alcance potencial de 1,69 bilhão de adultos maiores de 18 anos por mês) e WeChat (1,37 bilhão de usuários ativos mensais), que lideram o *ranking* global de presença digital.

A predominância do Twitter nos estudos científicos pode ser atribuída à sua política anteriormente aberta de compartilhamento de dados públicos, que facilitava a coleta e análise de informações pelos pesquisadores. No entanto, é importante destacar que as políticas atuais de compartilhamento de dados do Twitter, agora denominado X, tornaram-

se mais restritivas. Sob a nova administração, o acesso a esses dados passou a ser cobrado, o que representa um desafio significativo para novos estudos. Essa mudança alinha o X às diretrizes mais rigorosas de outras redes sociais em relação ao acesso e uso de dados, o que pode limitar consideravelmente sua aplicação em pesquisas acadêmicas futuras.

Com relação a **P4. Quais análises/tratamentos são realizados nos dados?** foi observado que a comunidade científica na maioria dos trabalhos fizeram análises e tratamentos convergentes, mostrando etapas bastante consolidadas a serem realizadas em análises de dados textuais de redes sociais. Dentre as análises e tratamentos mais empregados a remoção das *stopwords*[13] (28 trabalhos); a tokenização[14] (19); a remoção de *urls* (16); o *steeming*[15] (10); a remoção de sinais de pontuação (10); a conversão do texto para *lowercase* (8); a remoção de *tweets* duplicados (8); a normalização (5); a remoção de *hashtags* (5); a remoção de símbolos (4); a remoção de *retweets* (4); a lematização[16] (3); a representação do texto usando TF-IDF (3) e o uso de técnicas de balanceamento dos dados (3) foram os tratamentos mais utilizados nos trabalhos analisados.

Ao analisar os idiomas presentes nos conjuntos de dados utilizados nos estudos, para responder à questão **P5. Qual idioma dos dados textuais?** foi possível identificar os resultados apresentados na Figura 4.

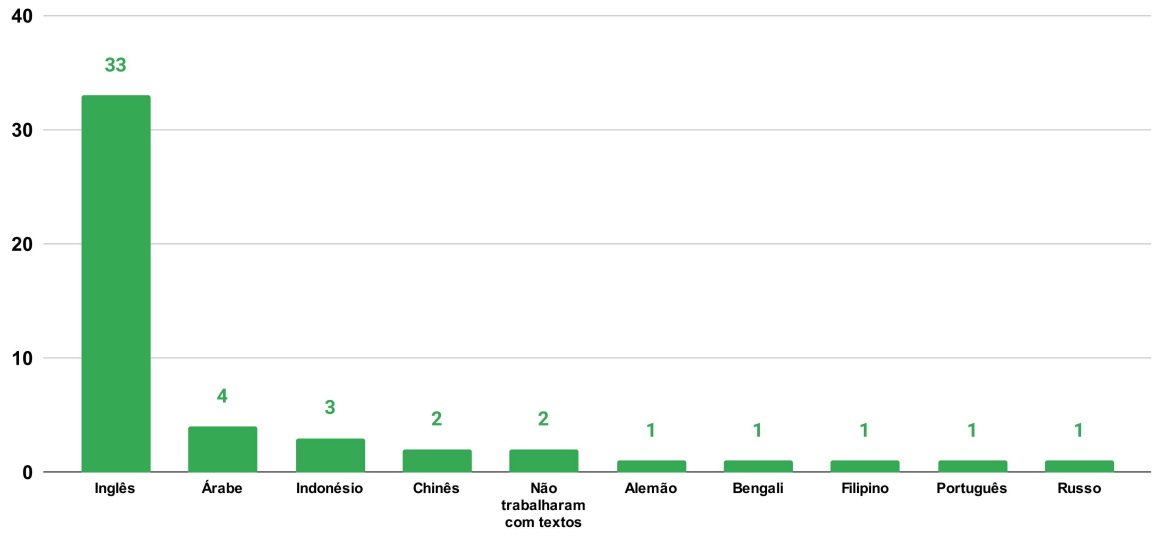


Figura 4 – Idiomas mais usados nos conjuntos de dados nos trabalhos analisados na revisão sistemática. **Fonte:** Autor.

Verificou-se que em 67,35% dos trabalhos o idioma usado foi o inglês, tornando-o o

[13] Palavras que se repetem muito em um texto e que não carregam consigo informação relevante como preposições e artigos.
[14] Processo que segmenta o texto a partir de um determinado critério como, por exemplo, o espaço em branco.
[15] Processo de redução de uma palavra à sua raiz.
[16] Processo de deflexionar uma palavra para determinar o seu lema. Por exemplo, as palavras correr, correria, corrida são todas formas do mesmo lema: correr.

mais utilizado. Em seguida os idiomas árabe (presente em 8,16% dos estudos), indonésio (6,12%), chinês (4,08%), não trabalharam com textos (4,08%), alemão, bengali, filipino, português e russo com (2,4%) com a mesma proporção de uso. Deste resultado é possível verificar a enorme representatividade do idioma inglês. Esse resultado corrobora com o que é descrito por EBERHARD DAVID M.; FENNIG (2022) em que o idioma inglês está no topo dos 10 idiomas mais falados no mundo, seguido do mandarim, hindi, espanhol, francês, árabe padrão, bengali, russo, português e urdu, justificando os resultados encontrados.

Um ponto notável é a presença de apenas um estudo que utilizou o idioma português, conduzido pela equipe do Observatório do Consumidor e descrito por NOGUEIRA et al. (2024). Essa baixa representatividade pode estar associada aos descritores de busca, em que os termos utilizados estavam em inglês, limitando assim a inclusão de trabalhos em outros idiomas. Apesar dessa limitação, estudos em diversos outros idiomas foram analisados, evidenciando a necessidade e a importância de incentivar pesquisas que utilizem o português. A ampliação de estudos nesse idioma contribuirá para o fortalecimento, o desenvolvimento e a melhoria de técnicas voltadas para a análise de textos em português, tanto nas redes sociais quanto em outros contextos.

Por fim, ao analisar os dados para responder à última questão secundária de pesquisa **P6. Quais são as métricas de avaliação?** identificou-se que as métricas mais utilizadas para avaliar os resultados foram a *precision* (20 trabalhos), *accuracy* (18), *F1-Score* (15), *recall* (15), curva ROC (5), curva AUC (4), KAV (4), *Mean Absoluto Error* (MAE) (4), matriz de confusão (4), *Mean Squared Error* (MSE) (4), *Root Mean Squared Error* (RMSE) (4), *Cohen Kappa* (1), correlação de Pearson (1), fórmula de *Slovin* (1), *G-mean* (1), *Log Likelihood Measurement* (1), *Mean Absolute Percentual Error* (MAPE) (1), média do sentimento de um determinado tópico (1), R^2 (1), relacionamento de variáveis (1), *Sum of Squared Error* (SSE) (1), TCR (1), *support* (1). Nove trabalhos não mencionaram as métricas de análise utilizadas.

Como em geral os contextos de aplicação envolvem problemas de classificação ou clusterização, as métricas mais utilizadas convergem entre cada um dos estudos e nos mostram as melhores métricas para serem aplicadas em cada situação.

2.2 Resposta da Questão Primária de Pesquisa

Após responder às questões secundárias de pesquisa, é possível abordar diretamente a questão principal que norteia este trabalho: **Quais métodos/ferramentas computacionais podem ser aplicados em dados textuais e de imagem oriundos das redes sociais para identificação de informações do mercado consumidor?**

Os resultados evidenciaram o uso predominante da Análise de Sentimentos, que se destacou como a abordagem mais utilizada para a compreensão de sentimentos e emoções em publicações *online*. Com a crescente disponibilidade de dados digitais e

a capacidade computacional robusta atualmente disponível, tornou-se viável analisar sentimentos expressos em milhares ou até milhões de publicações, tornando essa técnica um foco recorrente em pesquisas acadêmicas e aplicações práticas.

Além disso, a análise textual e a modelagem de tópicos também figuraram entre os métodos mais implementados. Estas técnicas permitem uma compreensão mais profunda dos contextos de aplicação, viabilizando a identificação de termos-chave e o agrupamento de temas relacionados, o que é essencial para explorar padrões e tendências no comportamento do consumidor.

Com relação aos tipos de dados, constatou-se que o uso de dados textuais é amplamente predominante em relação ao de imagens. Esse cenário reflete uma consolidação significativa de protocolos e boas práticas para o tratamento de dados textuais extraídos de redes sociais.

Entre as técnicas consideradas indispensáveis para lidar com esse tipo de dado estão a remoção de *stopwords*, remoção de URLs, tokenização, *stemming*, remoção de sinais de pontuação e conversão dos textos para *lowercase*. Essas etapas são fundamentais para a padronização, limpeza e preparação dos textos, facilitando o processamento e a análise subsequente. A adoção sistemática dessas técnicas reflete o amadurecimento das metodologias aplicadas em pesquisas que utilizam dados de redes sociais, garantindo maior precisão e qualidade nas análises realizadas.

Em diversos trabalhos, os autores precisaram criar seus próprios *datasets* para o treinamento de modelos, devido às características específicas de cada estudo. No entanto, um aspecto crítico que merece atenção é o balanceamento de classes. Muitos estudos analisados não relataram nem demonstraram o uso de técnicas para tratar o desbalanceamento entre classes nos conjuntos de dados utilizados.

Entre os trabalhos revisados, destacam-se os de JONATHAN et al. (2020) e NOGUEIRA et al. (2024), que foram os únicos a mencionar explicitamente o uso de técnicas como SMOTE, Tomek e SMOTE-Tomek para lidar com o desbalanceamento de classes. Esses métodos desempenham um papel fundamental ao ajustar a proporção de classes em conjuntos de dados, contribuindo diretamente para a obtenção de resultados mais consistentes e precisos.

A ausência generalizada de relatos sobre o uso de técnicas de balanceamento indica uma lacuna importante que precisa ser explorada. Estudos futuros poderiam focar na avaliação do impacto de diferentes abordagens de balanceamento, como *oversampling*, *undersampling* e algoritmos híbridos, para gerar conjuntos de dados mais representativos e melhorar o desempenho de modelos aplicados. Esse esforço é essencial para garantir que os resultados obtidos reflitam com maior fidelidade as características reais dos fenômenos estudados.

Em relação às principais redes sociais utilizadas nos trabalhos analisados, verificou-

se que o Twitter, foi a plataforma mais empregada na condução de pesquisas, seguido pelo Facebook e pelo YouTube. O destaque do Twitter estava diretamente relacionado à sua política de compartilhamento de dados públicos, historicamente mais aberta em comparação com outras redes sociais. Entretanto, com as mudanças na administração da plataforma, as políticas de acesso a dados tornaram-se significativamente mais restritivas, incluindo a imposição de cobranças pelo acesso a informações que antes eram públicas. Essa alteração representa um desafio para pesquisadores, podendo limitar a coleta, o processamento e a análise de dados dessa rede social em estudos futuros, além de abrir espaço para que outras plataformas assumam maior relevância no campo acadêmico.

No que diz respeito às técnicas de processamento de linguagem natural aplicadas, não foram identificados trabalhos que explorassem as relações morfo-sintáticas presentes nos textos analisados. Essa lacuna sugere um caminho promissor para futuras pesquisas na área. O estudo das classes gramaticais e de suas relações pode oferecer informações como a identificação de ações realizadas pelos usuários a partir dos verbos, bem como a descrição de características e qualidades de produtos com base no uso de adjetivos. Explorar essas relações pode enriquecer a análise dos dados textuais, proporcionando uma compreensão mais profunda e detalhada dos conteúdos presentes nas redes sociais, além de ampliar as possibilidades de aplicação prática no mercado consumidor.

Em relação ao idioma dos conjuntos de dados analisados, o inglês apresentou a maior representatividade. Esse resultado reflete duas perspectivas importantes. A primeira é que, por ser amplamente utilizado como idioma universal, o inglês facilita a realização de pesquisas com maior alcance, além de possuir um volume significativo de estudos contextualizados e validados pela comunidade científica.

Por outro lado, a baixa representatividade de idiomas latino-americanos, como o português e o espanhol, destaca uma lacuna significativa na literatura. No caso do português, apenas um estudo foi identificado no escopo das palavras-chave utilizadas nesta pesquisa. Esse dado evidencia a necessidade e a oportunidade de promover mais investigações acadêmicas que explorem textos nesse idioma, fomentando o desenvolvimento de metodologias e técnicas específicas que atendam às peculiaridades linguísticas e culturais destes países.

O uso de imagens provenientes de redes sociais representa um campo de grande potencial a ser explorado. Em plataformas como Instagram, Facebook, Flickr e Twitter, as imagens desempenham um papel central, seja em postagens de usuários, seja em fotografias de perfil. A análise desses elementos visuais abre diversas possibilidades de pesquisa e aplicação. Com o uso de algoritmos avançados e modelos preexistentes, é possível empregar técnicas de reconhecimento facial, identificar entidades presentes nas imagens e detectar padrões associados a tendências de consumo. Essas análises permitem o cruzamento de dados visuais com outros contextos, proporcionando informações sobre comportamento,

preferências e interações sociais.

Contudo, para identificar informações do mercado consumidor, conclui-se que a aplicação de técnicas como análise de sentimentos, análise textual ou modelagem de tópicos são muito promissoras. Essas metodologias permitem extrair características textuais, identificar a polaridade de sentimentos e interpretar as emoções expressas pelos usuários de redes sociais em relação a um tema específico. A integração dessas técnicas possibilita uma compreensão aprofundada dos comportamentos e hábitos dos usuários, fornecendo dados valiosos que podem ser utilizados de maneira estratégica pelo mercado. Além disso, elas oferecem oportunidades para identificar padrões inéditos ou replicar abordagens semelhantes aos observados nos estudos analisados nesta revisão. Isso evidencia o potencial dessas abordagens para inovar na exploração de dados de redes sociais, contribuindo significativamente para a formulação de estratégias direcionadas e informadas.

3 REFERENCIAL TEÓRICO

Este capítulo apresenta, em detalhes, os conceitos fundamentais que compõem a base metodológica deste trabalho, abordando os aspectos teóricos relacionados aos processos do Observatório do Consumidor. Além disso, são descritas as principais referências e definições de termos que serão mencionados nos capítulos subsequentes, com o objetivo de proporcionar uma compreensão clara e fundamentada do arcabouço científico que sustenta este estudo.

3.1 Linguística Computacional

A linguística computacional é uma área interdisciplinar que integra conhecimentos de linguística, ciência da computação e inteligência artificial para desenvolver sistemas capazes de processar, compreender e gerar linguagem humana. De acordo com VIEIRA; LIMA (2001), a linguística computacional pode ser definida como "*a área do conhecimento que explora as relações entre linguística e informática, viabilizando a criação de sistemas com capacidade de reconhecer e produzir informações expressas em linguagem natural*".

O foco central dessa área está na automação de tarefas linguísticas, tais como análise sintática, tradução automática de idiomas, extração de informações, classificação e interpretação de textos. Dessa forma, o objetivo das próximas subseções é o de apresentar as principais técnicas utilizadas no processamento de textos, introduzindo conceitos e termos fundamentais.

3.1.1 Processamento de linguagem natural

A necessidade de ensinar computadores a executar tarefas humanas impulsionou o desenvolvimento e aprimoramento de diversas áreas da ciência, com destaque para a computação. Entre essas áreas, destaca-se o Processamento de Linguagem Natural (PLN), cujo principal objetivo é identificar e extrair representações e significados de textos em linguagem natural[1] (INDURKHYA; DAMERAU, 2010).

Um dos maiores desafios do PLN reside na tradução da linguagem humana para um formato que os computadores possam compreender e processar. Esse desafio envolve lidar com as nuances, ambiguidades e complexidades inerentes à linguagem natural, tornando o desenvolvimento de sistemas eficientes um campo de estudo essencial na inteligência artificial e na linguística computacional.

Diversos fatores tornam os trabalhos que utilizam PLN desafiadores. NOGUEIRA (2021) destaca que a natureza dos textos, sejam eles formais (no padrão culto da língua) ou informais (como os presentes em redes sociais), pode impactar significativamente os

[1] Linguagem utilizada na comunicação humana, seja falada ou escrita, como português, inglês ou espanhol, entre outras.

resultados. Textos formais, por seguirem um padrão de formatação mais estruturado, tendem a facilitar o processamento.

Por outro lado, nos dados de redes sociais, foco deste trabalho, prevalece a linguagem informal, caracterizada por palavras abreviadas, erros ortográficos, uso de *emojis*, além de ironia e sarcasmo. Esses elementos introduzem ruídos que dificultam o aprendizado de máquina, exigindo a aplicação de técnicas de tratamento e normalização dos dados para melhorar a qualidade do processamento (NOGUEIRA, 2021).

LIU; ZHANG (2012) destacam em seu estudo que a análise da linguagem a partir de seu significado impulsionou o desenvolvimento de diversos sistemas baseados PLN. Ao reconhecer que um texto vai além de uma simples sequência de caracteres, possuindo uma estrutura hierárquica, foi possível criar aplicações amplamente utilizadas atualmente. Entre elas, destacam-se os corretores gramaticais, a transcrição da linguagem falada em texto, a tradução automática, e até mesmo a identificação de emoções e sentimentos presentes em textos, por meio da análise de sentimentos, também conhecida como mineração de opinião.

3.1.1.1 Pré-processamento de dados textuais

O fluxo do PLN consiste em uma série de etapas de tratamento responsáveis por preparar e padronizar os dados textuais, tornando-os adequados para análise e uso em aplicações. A Figura 5 apresenta algumas etapas que envolvem as tarefas de pré-processamento de textos.

Figura 5 – 7 etapas de pré-processamentos textuais. **Fonte:** Autor.

3.1.1.1.1 Limpeza e normalização

Devido à diversidade na forma como um texto pode ser escrito, cada contexto ou ambiente apresenta um padrão de escrita específico. A formalidade ou não de um texto influencia diretamente em sua estrutura, levando as pessoas a utilizarem diferentes estilos e convenções. Por exemplo, há textos em que todas as palavras podem estar em letras maiúsculas, outros que podem alternar entre maiúsculas e minúsculas, e aqueles escritos exclusivamente em letras minúsculas.

Diante dessa variação, a normalização desempenha um papel essencial no PLN, visando padronizar o texto e facilitar seu tratamento e análise (CHEN et al., 2024). Para exemplificação das etapas de processamento mencionadas na Figura 5, a Figura 6 mostra um exemplo de frase que necessita de vários procedimentos de tratamento.

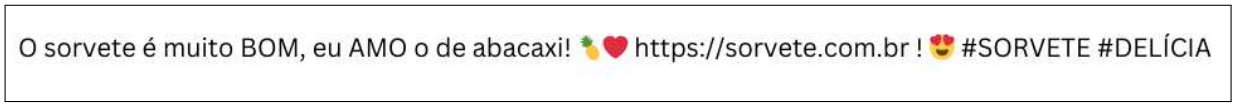


Figura 6 – Exemplo de frase para demonstrar as etapas do pré-processamento de texto. **Fonte:** Autor.

A Figura 6 apresenta um texto com características típicas de dados de redes sociais que evidenciam a necessidade de realizar a limpeza e normalização dos dados para facilitar sua análise. O primeiro passo nesse processo é a conversão de letras maiúsculas (*uppercase*) para minúsculas (*lowercase*) por meio da técnica de *lowercasing*, que garante maior uniformidade no texto, reduz ambiguidades e facilita as etapas subsequentes.

Em seguida, ocorre a remoção de ruídos — elementos que dificultam a análise ou não contribuem diretamente para os objetivos do processamento textual. Entre eles estão os *emojis* e *emoticons*, usados para expressar emoções, *hashtags* e os *links* de sites, frequentemente utilizados para compartilhar conteúdos multimídia.

Outros ruídos incluem sinais de pontuação, símbolos, números e, em certos casos, palavras comuns (*stopwords*) sem valor semântico relevante. A remoção desses componentes é essencial para "limpar" o texto, reduzindo interferências indesejadas e aprimorando os resultados da análise. A Figura 7, exibe o resultado da transformação do texto para *lowercase* e remoção de *emojis*, *links* de sites, sinais de pontuação e *hashtags*.

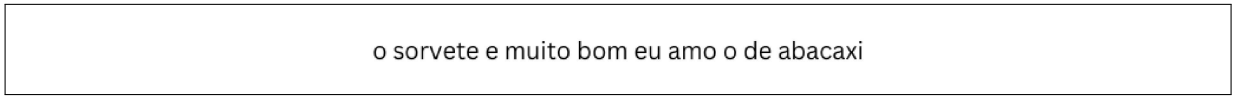


Figura 7 – Resultado da aplicação da limpeza e normalização da frase de exemplo. **Fonte:** Autor.

3.1.1.1.2 Tokenização

A segunda etapa, a tokenização, é conhecida como o processo de segmentar as palavras de um texto, ou seja, é a responsável por quebrar uma sequência de caracteres, o objetivo desta técnica é identificar a localização em que uma palavra termina dando início a outra (PALMER, 2010). O termo palavra, em linguística computacional, é frequentemente chamado de *token*.

Um dos separadores mais utilizados em línguas que podem ser segmentadas como, por exemplo, inglês, português, espanhol é o espaço em branco, em que se descarta este caractere mantendo somente os *tokens*. Outros delimitadores que podem ser utilizados são vírgulas, pontos e vírgulas, sinais de pontuação, dentre outros. Em linguagens que não podem ser segmentadas, como o chinês e o tailandês, as palavras são escritas sem nenhuma indicação de término. Em linguagens deste tipo, este processo necessita de informações léxicas e morfológicas adicionais. A Figura 8 exibe o resultado da aplicação da tokenização na frase de exemplo normalizada exposta na Figura 7.

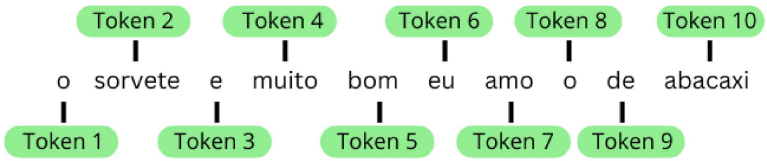


Figura 8 – Resultado da aplicação da tokenização na frase de exemplo. **Fonte:** Autor.

3.1.1.1.3 Remoção de *stopwords*

Outro passo importante, amplamente destacado na revisão sistemática da literatura apresentada na Seção 2, é a remoção das *stopwords*. Como observado, a comunidade científica, ao trabalhar com textos, têm utilizado regularmente técnicas para eliminar caracteres e palavras que podem dificultar o processamento e o aprendizado de modelos computacionais.

No tratamento de dados textuais, as *stopwords* são palavras que, por serem muito comuns, têm pouco ou nenhum impacto no significado, ou no contexto principal de um texto (CHEN et al., 2024). Essas palavras geralmente incluem artigos, preposições, pronomes e conjunções, como “o”, “a”, “‘de”, “para”, “e” e “em”. Sua remoção é uma prática comum para “limpar” o texto, tornando-o mais eficiente e direto para o processamento computacional, como destacado em SOLKA (2008).

Neste contexto, ao aplicar essa técnica à frase tokenizada apresentada na Figura 8, observa-se que os *tokens* 1, 3, 8 e 9 são removidos. O *token* 1 e o 8 correspondem a

artigos, o 3 é uma conjunção, e o 9 é uma preposição. Após a remoção desses elementos, a frase resultante, já tratada, é exibida na Figura 9.

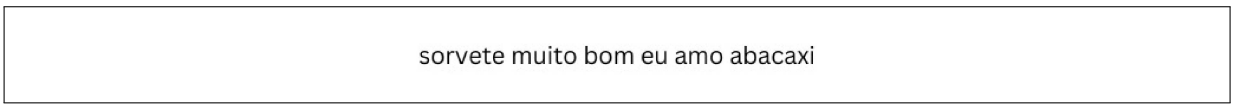


Figura 9 – Resultado da frase processada. **Fonte:** Autor.

3.1.1.1.4 Lematização e stemização

Conforme destacado na revisão sistemática apresentada na Seção 2, a lematização e a stemização são dois processos utilizados em estudos de PLN. Ambos procuram manipular as palavras de um texto para facilitar a análise textual. A lematização consiste em reduzir as formas flexionadas de uma palavra à sua forma base, garantindo que essa forma reduzida seja válida no idioma (CHEN et al., 2024). Essa forma base é conhecida como lema. Por exemplo, as palavras "correr", "correndo" e "correria" são flexões de "correr", que é o lema. A lematização é crucial porque reduz a diversidade de formas lexicais, permitindo que diferentes flexões de uma palavra sejam tratadas como uma única unidade representativa. Isso contribui para a normalização do texto e melhora a consistência nas análises linguísticas e computacionais.

Já a stemização também busca reduzir as palavras a uma forma simplificada, mas com foco em identificar a raiz ou o núcleo comum das palavras, chamado de *stem* (tronco) (CHEN et al., 2024). Contudo, ao contrário da lematização, a stemização não assegura que a forma reduzida pertença ao idioma. Por exemplo, as palavras "pata" e "pato" podem ser reduzidas ao *stem* "pat", que não corresponde a uma palavra válida no idioma. Apesar de suas diferenças, ambos os processos são ferramentas importantes no pré-processamento de dados textuais, dependendo do objetivo da análise.

3.1.1.1.5 TF-IDF - *Term Frequency-Inverse Document Frequency*

O TF-IDF é uma técnica amplamente utilizada em PLN para avaliar a relevância de uma palavra em relação a um documento em um conjunto de documentos (*corpus*) (HAVRLANT; KREINOVICH, 2017). Ele combina duas métricas: a frequência do termo (TF), que mede o número de vezes que uma palavra aparece em um documento, e a frequência inversa do documento (IDF), que pondera a raridade da palavra em todo o *corpus*. Palavras comuns, como artigos e preposições, recebem pontuações mais baixas, enquanto termos mais específicos e representativos têm valores mais altos, tornando o TF-IDF uma ferramenta eficaz para tarefas como classificação de texto, busca de informações e extração de palavras-chave.

3.1.1.1.6 *Part-of-speech* (POS) *Tagging*

O processo de *Part-of-Speech Tagging* (na tradução para o português, "etiquetagem de parte da fala" ou "marcação de classe gramatical") também é muito utilizado no PLN e está diretamente relacionado à identificação das classes gramaticais (CHEN et al., 2024). O objetivo principal desse processo é atribuir a cada palavra de uma frase sua respectiva função morfológica[2], dependendo do contexto em que é empregada (ROTH; ZELENKO, 1998).

Esse processo é fundamental em diversas tarefas do PLN, pois permite uma compreensão mais estruturada da linguagem natural. Por exemplo, a palavra "corre" pode ser identificada como um verbo na frase "Ela corre todos os dias" ou como um substantivo em "O corre da cidade grande". A etiquetagem de *POS* analisa esses contextos e classifica corretamente as palavras, auxiliando em diversas aplicações. A Figura 10 ilustra um exemplo prático desse processo, demonstrando como as palavras de uma frase são anotadas com suas respectivas classes gramaticais.

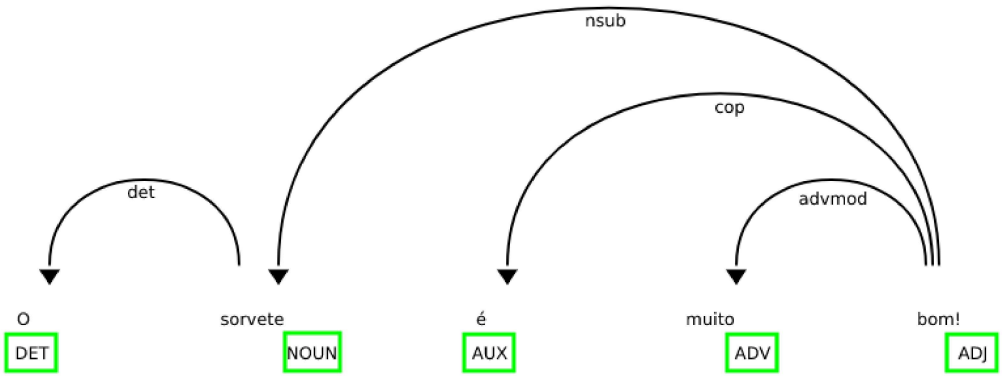


Figura 10 – Sentença anotada utilizando o processo *Part-of-Speech Tagging* utilizando a biblioteca spaCy (HONNIBAL; MONTANI, 2017). Em verde é destacado as *tags* anotadas referentes aos *tokens*. **Fonte:** Autor.

Abaixo de cada uma das palavras da frase "O sorvete é muito bom!" estão localizadas as chamadas *POS tags*. Neste simples exemplo é possível verificar a presença de cinco *tags*: DET (artigo), NOUN (substantivo), AUX (verbo auxiliar), ADV (advérbio) e ADJ (adjetivo). Entretanto, existem várias outras *tags*[3] que estarão presentes em outros contextos e frases.

[2] Na Língua Portuguesa, a análise morfológica refere-se ao estudo das palavras individualmente, considerando seus aspectos gramaticais, como gênero, número, tempo verbal, entre outros, sem levar em conta sua relação sintática com outras palavras.

[3] *Link* com diversas *tags* denominadas universais https://universaldependencies.org. Acesso em: 13 fev. 2025.

3.1.1.1.7 *Dependency Parsing*

A análise de dependência[4] (do inglês, *Dependency Parsing* - DP) é uma tarefa essencial no PLN, voltada para a captura de informações sintáticas em sentenças por meio das relações de dependência entre palavras (HAN et al., 2020). Esse processo consiste em extrair o grafo de dependência de uma frase, representando sua estrutura gramatical de maneira hierárquica. A análise estabelece as relações entre cada *token* e seus dependentes, destacando como as palavras de uma sentença se conectam e desempenham suas funções no contexto.

O elemento central dessa análise é a cabeça (*head*) – também chamada de raiz –, que não possui dependências. Segundo KULMIZEV (2023) em geral, a cabeça é o verbo principal da frase, pois ela desempenha o papel central na estrutura gramatical. A estrutura do grafo gerado pela análise de dependência é representada por um grafo dirigido, em que os nós correspondem aos *tokens* (as palavras ou elementos da frase), e as arestas (ou setas) representam as relações gramaticais entre os *tokens*.

Essa abordagem possibilita identificar o papel gramatical desempenhado por cada *token* no texto, como sujeito, objeto, modificadores, entre outros elementos, além de compreender as interações entre eles. A Figura 11 ilustra essa análise, destacando os elementos do grafo de dependência e as relações sintáticas estabelecidas entre as palavras.

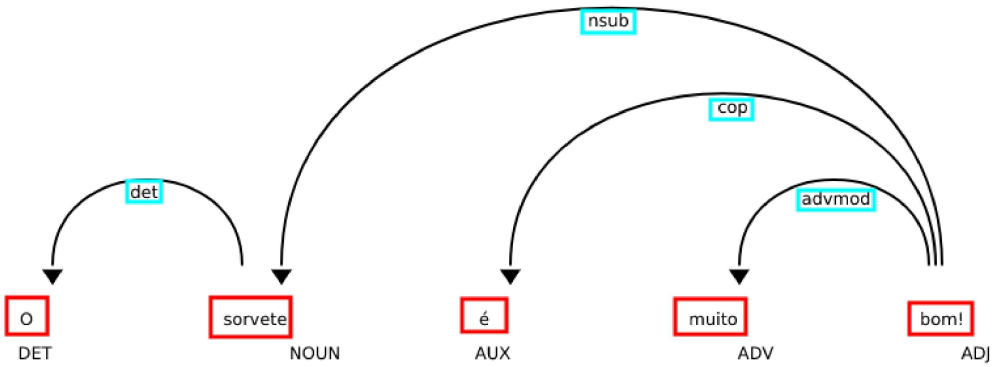


Figura 11 – Análise de dependência anotada utilizando o processo *Dependency Parsing* utilizando a biblioteca spaCy (HONNIBAL; MONTANI, 2017). Em vermelho é destacado os *tokens* da frase e em ciano as *tags* anotadas referentes às relações de dependência dos *tokens*. **Fonte:** Autor.

[4] Também conhecida como análise sintática na Língua Portuguesa, refere-se ao estudo das palavras de uma oração e das funções que desempenham em relação umas às outras.

3.2 Análise de Sentimentos

Após definir e exemplificar termos relacionados às etapas de pré-processamento de dados textuais, esta seção explora as possibilidades de análise que podem ser realizadas após o tratamento e processamento desses dados. Conforme observado no capítulo 2, existem várias abordagens possíveis. No entanto, a comunidade científica, como destacado na revisão, concentra-se majoritariamente na aplicação da análise de sentimentos, especialmente em estudos que envolvem dados textuais coletados de redes sociais.

A análise de sentimentos, conforme definida por LIU (2012), é uma área de estudo voltada para a extração de informações a partir de opiniões expressas sobre um determinado tema, seja em textos, imagens, vídeos ou outros formatos. KANG; PARK (2014) destacam que essa técnica também pode ser utilizada para avaliar a satisfação do usuário em relação a produtos ou serviços, permitindo que empresas identifiquem e solucionem pontos fracos em seus empreendimentos de maneira mais eficaz.

Em essência, a análise de sentimentos é uma abordagem que identifica emoções e sentimentos manifestados em dados relacionados a um contexto específico. Com o auxílio de recursos computacionais, essa técnica automatizada possibilita a obtenção de visões globais sobre um tema de forma rápida e eficiente, superando a subjetividade e a limitação de tempo associadas a análises manuais realizadas por humanos.

No capítulo 2, constatou-se que a análise de sentimentos foi aplicada em 27,16% dos trabalhos analisados, abrangendo 22 contextos distintos de aplicação, o que evidencia a ampla diversidade de uso dessa técnica. No entanto, para alcançar resultados satisfatórios, é fundamental superar determinados desafios que são cruciais para a obtenção de análises precisas e eficazes. O primeiro está relacionado com o conjunto de dados de treinamento, também chamado de *dataset*. Dependendo do contexto de aplicação, a disponibilidade de dados pode ser substancialmente limitada, o que impacta diretamente a capacidade de aprendizado do modelo. Outro fator relevante diz respeito às classes às quais cada amostra pertence. Além disso, surge a questão de como rotular esses dados, especialmente quando se lida com grandes volumes de informações. A representatividade de dados em cada classe, ou seja, o balanceamento entre as classes, também exerce um papel crucial. Assim, as próximas subseções exploram em maior detalhe as estratégias atualmente empregadas e potenciais abordagens para superar esses desafios.

3.2.1 Criação de *datasets* de treinamento

Um requisito essencial para tarefas que utilizam técnicas tradicionais de aprendizado de máquina supervisionados é a obtenção de dados que representem de forma adequada o contexto de aplicação de interesse e apresentem qualidade suficiente uma vez que o desempenho do modelo é diretamente influenciado pela qualidade dos dados de treinamento (AZATISOFTWARE, 2019). Em problemas de classificação, como na

análise de sentimentos, alcançar uma representação equilibrada entre as classes estimadas pelo modelo é fundamental para obter resultados mais robustos e confiáveis. Na análise dos trabalhos apresentados na revisão apresentada no capítulo 2, observou-se que os autores frequentemente utilizavam dois ou três tipos de classes. No primeiro cenário, foram consideradas apenas as classes “negativa” e “positiva”, enquanto no segundo, além dessas, foi incluída a classe “neutra”.

Diversas técnicas podem ser empregadas para rotular ou etiquetar dados em classes específicas, destacando-se três principais abordagens: manual, semi-automática e automática.

3.2.1.1 Estratégia manual de rotulação

Uma das abordagens mais comuns para rotulação de dados envolve especialistas de um determinado contexto que analisam manualmente cada instância do conjunto de dados, atribuindo-lhe um rótulo de acordo com critérios previamente definidos (IBM, 2024). Outra estratégia amplamente utilizada é o *crowdsourcing*, no qual diversas pessoas, por meio de uma plataforma, fornecem os rótulos que consideram mais adequados para cada instância (FREDRIKSSON et al., 2020). Plataformas populares para essa tarefa incluem a *Amazon Mechanical Turk* e a *Lionbridge AI* (HACKERNOON, 2020).

Em um dos trabalhos analisados na Seção 2, AL-YASIRI; AL-AZAWEI (2019) descrevem a construção de um *dataset* com 4.762 *tweets*, dos quais 2.439 foram classificados como positivos e 2.323 como negativos utilizando o processo de rotulação manual. Para cada *tweet*, um rótulo (positivo ou negativo) foi atribuído com base no entendimento de uma pessoa ao ler o texto completo.

Embora esse tipo de rotulação seja comum em muitos estudos, ele pode se tornar mecânico e cansativo, especialmente em grandes volumes de dados, podendo introduzir ruídos que afetam a qualidade do modelo. De acordo com PEREZ; WANG (2017), quanto maior a quantidade de dados fornecidos a um algoritmo de aprendizado de máquina, maiores as chances de obter um modelo eficiente. Isso implica que, em situações que exigem grandes volumes de dados para treinamento, a rotulação manual pode ser ineficaz e inadequada.

3.2.1.2 Estratégias semi-automáticas de rotulação

As estratégias semi-automáticas buscam combinar as abordagens manual e automática, aproveitando as vantagens de ambas. Nesse método, os modelos selecionam amostras mais informativas de um grande conjunto de dados não rotulados e, em seguida, consultam um oráculo (por exemplo, um anotador humano) para atribuir rótulos de forma iterativa (EL-HASNONY et al., 2022). Durante esse processo, o modelo aprende observando as de-

cisões de rotulação realizadas pelos humanos, utilizando esse aprendizado para recomendar rótulos e automatizar funções básicas na interface de rotulação (DESMOND et al., 2021).

As principais vantagens dessa abordagem incluem a redução significativa na quantidade de dados que precisam ser rotulados manualmente, além da capacidade de alinhar os rótulos com as expectativas do usuário final da aplicação. No entanto, a estratégia apresenta algumas limitações, como a necessidade de integração eficiente entre os processos de rotulação manual e automática. Além disso, pode ser um processo lento e exigir esforço contínuo dos usuários envolvidos na rotulação.

3.2.1.3 Estratégias automáticas de rotulação

A rotulação automática visa superar as limitações da abordagem manual, tornando o processo mais escalável e viável para grandes volumes de dados. Essa estratégia envolve diversas técnicas, como a rotulação baseada em regras, a utilização de modelos pré-treinados e a transferência de rótulos por meio de *Transfer Learning*. A técnica de rotulação baseada em regras classifica automaticamente os dados que apresentam características específicas de interesse (CAVALCANTE; BARBOSA, 2017). A abordagem com modelos pré-treinados, por sua vez, utiliza grandes conjuntos de dados para treinar os modelos, tornando-os altamente eficientes na resolução de problemas complexos (DEVLIN et al., 2019). Já o *Transfer Learning* transfere rótulos de um conjunto de dados existente para outro com características semelhantes, aproveitando o conhecimento prévio adquirido para aprimorar a precisão da rotulação em novos contextos (ZHUANG et al., 2020).

CAVALCANTE; BARBOSA (2017) rotularam dados de *posts* do X, antigo Twitter, baseando-se na rotulação baseada em regras. Publicações que continham *emojis* “:)” ou “:-)” eram classificados como positivos, enquanto os que continham “:(” ou “:-(” foram classificados como negativos. KOULOUMPIS et al. (2011) relataram que muitos pesquisadores utilizam essa abordagem para determinar o sentimento do conjunto de dados de treinamento. Usando essa metodologia, os autores coletaram 75.828 *tweets*, com 50% dos dados pertencendo à classe positiva e 50% à classe negativa.

PINTO; ROCIO (2019) conduziram outro estudo utilizando uma API para rotulação automática de dados e construção de conjuntos de dados. Eles empregaram quatro APIs: *Amazon Comprehend*, *Google Natural Language*, *IBM Watson Natural Language Understanding* e *Microsoft Text Analytics*. Os autores submeteram dados coletados da plataforma *Moodle* a cada ferramenta para recuperar as pontuações de sentimento de cada texto. Eles combinaram as pontuações obtidas de cada API[5] para construir o conjunto de dados. O objetivo dos autores era investigar a relação entre a polaridade do sentimento expresso em intervenções online e o desempenho acadêmico dos estudantes da Universidade Aberta de Portugal.

[5] Sigla para *Application Programming Interface*.

As abordagens de rotulação automática apresentam várias vantagens, como rapidez, escalabilidade e alta eficiência, sendo especialmente eficazes em grandes volumes de dados e tarefas complexas. Elas também se destacam pela facilidade de implementação em problemas bem definidos e pela redução significativa do esforço necessário para a rotulação inicial. No entanto, essas abordagens possuem algumas desvantagens, principalmente em relação à precisão, já que, ao contrário da análise minuciosa realizada por especialistas, não oferecem total controle sobre a qualidade da classificação. Além disso, os resultados obtidos dependem diretamente da qualidade do modelo utilizado, o que pode resultar em imprecisões, especialmente se houver grandes variações entre os contextos de aplicação. Por essa razão, o uso de ferramentas confiáveis, como as mencionadas no estudo de PINTO; ROCIO (2019), são cruciais para garantir a eficácia da rotulação automática.

3.2.2 Abordagens para mitigar o desbalanceamento de dados

Em todas as estratégias de rotulação de dados, um fator crucial, já mencionado anteriormente, é a representatividade de cada classe. JONATHAN et al. (2020) destacam que garantir uma boa representatividade entre as classes é fundamental para alcançar resultados satisfatórios. Muitas abordagens foram desenvolvidas para enfrentar o desafio apresentado por conjuntos de dados desbalanceados e aprimorar o desempenho dos modelos de aprendizado de máquina (NOGUEIRA et al., 2024).

Essas abordagens englobam técnicas de subamostragem (*under-sampling*) e superamostragem (*over-sampling*), cada uma adaptada para reequilibrar a distribuição das classes dentro do conjunto de dados, melhorando assim a robustez e a precisão dos modelos preditivos. Nas próximas subseções, serão abordadas seis dessas técnicas, a saber: (i) *NearMiss*, (ii) *SMOTE*, (iii) *ADASYN*, (iv) *SMOTE-ENN*, (v) *SMOTETomek* e (vi) *Back Translation*.

3.2.2.1 *NearMiss*

A técnica *NearMiss* (ZHANG; MANI, 2003) é uma abordagem de *under-sampling* que inclui três variantes baseadas no algoritmo *K-Nearest Neighbors*. Essas variantes são: (i) *NearMiss-1*, que seleciona as amostras da classe majoritária com a menor distância média para os k-vizinhos mais próximos da classe minoritária; (ii) *NearMiss-2*, que escolhe as amostras da classe majoritária com a menor distância média para os k-vizinhos mais distantes da classe minoritária; e (iii) *NearMiss-3*, que, para cada amostra da classe minoritária mais próxima, seleciona um número pré-determinado de amostras da classe majoritária.

3.2.2.2 SMOTE - *Synthetic Minority Oversampling Technique*

Um dos métodos de *over-sampling* mais conhecidos, desenvolvido por CHAWLA et al. (2002), é o *SMOTE*. Para garantir que as amostras geradas sejam distintas da classe minoritária original, essa técnica cria amostras sintéticas com base na distância entre cada ponto de dados e seus vizinhos mais próximos da classe minoritária. O processo de geração de amostras sintéticas pode ser resumido nos seguintes quatro passos:

- Seleciona-se dados aleatórios da classe minoritária;
- Calcule a distância euclidiana entre os dados selecionados e seus k-vizinhos mais próximos;
- Multiplique a diferença entre os dados e os vizinhos por um número aleatório entre 0 e 1, e adicione o resultado à classe minoritária como uma amostra sintética;
- Repita o procedimento até alcançar a proporção desejada da classe minoritária.

Este método é altamente eficaz, pois as amostras sintéticas geradas ficam relativamente próximas ao espaço de características da classe minoritária, preservando as propriedades dos dados originais.

3.2.2.3 *ADASYN - Adaptive Synthetic*

Esta também é uma técnica de *over-sampling*, semelhante à *SMOTE* mas com uma diferença significativa. Enquanto o *SMOTE* gera amostras sintéticas com base em vizinhos próximos, o *ADASYN* utiliza um classificador *k-Nearest Neighbors* (k-NN) para gerar dados sintéticos que são semelhantes às amostras originais classificadas erroneamente (HE et al., 2008). A função de decisão durante o treinamento se difere entre os métodos devido à implementação básica do *SMOTE* não conseguir distinguir amostras fáceis das difíceis, por meio da regra dos vizinhos mais próximos. A seguir estão os passos realizados por este algoritmo:

- Determine o número total de amostras que serão produzidas;
- Para cada amostra minoritária, identifique os seus k-vizinhos mais próximos e calcule a proporção entre k e o número de vizinhos da classe majoritária mais próximos;
- Calcule o número de amostras a serem geradas a partir de cada amostra minoritária, proporcional à razão calculada no passo anterior;
- Por fim, utilize interpolação entre a amostra minoritária e seus vizinhos mais próximos para gerar o número necessário de amostras sintéticas.

Essa abordagem visa melhorar a geração de amostras sintéticas, priorizando as amostras que são mais difíceis de classificar, o que pode resultar em uma maior eficiência no treinamento de modelos.

3.2.2.4 *SMOTE-ENN - SMOTE & Edited Nearest-Neighbours*

O *SMOTE & Edited Nearest-Neighbours* é um método que combina técnicas de *over-sampling* e *under-sampling*. Desenvolvido por BATISTA et al. (2004), este método faz a combinação do algoritmo *SMOTE* que gera amostras sintéticas para a classe minoritária com a capacidade *ENN* que elimina amostras cujo rótulo da classe se difere da classe da maioria de seus k-vizinhos mais próximos.

3.2.2.5 *SMOTE & TOMEK-Links* - SMOTETomek

Assim como o método *SMOTE & Nearest-Neighbours*, o *SMOTE & TOMEK-Links* também combina técnicas de *over-sampling* e *under-sampling*. Introduzido inicialmente por BATISTA et al. (2003), este método faz a combinação da capacidade do algoritmo *SMOTE* de criar dados sintéticos para a classe minoritária com a capacidade *Tomek-Links* que elimina os dados da classe majoritária classificadas como um *Tomek Link*.

Um *Tomek link* existe entre duas amostras a e b com rótulos diferentes se houver vizinhos mutuamente mais próximos, ou seja, se o vizinho mais próximo de a é b e vice-versa. O que acontece na prática é que ou remove-se todas as amostras pertencentes a um *Tomek link* ou apenas as amostras da classe majoritária pertencente a um *Tomek link*.

3.2.2.6 *Back Translation*

Outro método amplamente utilizado em tradução automática é o do *Back Translation* (SENNRICH et al., 2016; LAMPLE et al., 2018). Essa abordagem envolve o uso de um modelo de tradução pré-treinado, da língua alvo para a língua de origem, para gerar texto original ao realizar a tradução reversa do texto alvo. De acordo com SENNRICH et al. (2016), a inclusão de dados traduzidos de volta no conjunto de treinamento resulta em melhorias significativas no desempenho dos modelos de tradução automática. Para facilitar a comparação entre os métodos de balanceamento de dados, foi criada uma tabela comparativa, apresentada na Tabela 3.

Tabela 3 – Tabela comparativa das vantagens e desvantagens de cada método de balanceamento de dados utilizado neste estudo. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Método	Vantagens	Desvantagens	Rótulo
NearMiss (ZHANG; MANI, 2003)	- Eficiente para grandes conjuntos de dados - Preserva a estrutura do conjunto de dados - Pode melhorar a generalização do modelo	- Pode resultar na perda de informações importantes da classe majoritária - Sensível a *outliers*	A
SMOTE (CHAWLA et al., 2002)	- Eficiente para aumentar o tamanho do conjunto de dados - Gera amostras sintéticas diversas - Melhora a separação das classes em regiões densas dos dados	- Pode aumentar o risco de *overfitting* - Sensível a *outliers*	B
ADASYN (HE et al., 2008)	- Gera amostras sintéticas adaptativas específicas para instâncias minoritárias - Trata desequilíbrios de classe em diferentes regiões do espaço de características	- Pode ser intensivo computacionalmente em grandes conjuntos de dados - Sensível a *outliers*	B
SMOTE & Edited Nearest Neighbors (BATISTA et al., 2004)	- Reduz o risco de gerar amostras sintéticas não representativas - Melhora a qualidade das amostras geradas	- Pode não remover completamente amostras ruidosas - Sensível a desequilíbrios extremos	C
SMOTE & TOMEK-Links (BATISTA et al., 2003)	- Melhora a separabilidade entre as classes - Pode ajudar a reduzir o desequilíbrio do conjunto de dados	- Pode remover amostras relevantes da classe majoritária - Sensível a amostras ruidosas	C
Back Translation (SENNRICH et al., 2016); (LAMPLE et al., 2018)	- Aumenta a quantidade de dados disponíveis para treinamento - Melhora o desempenho dos modelos de tradução automática	- Requer um modelo de tradução pré-treinado de alta qualidade - Pode introduzir erros de tradução e distorcer o significado original do texto	D

Legenda: A - Aplica poda de dados. B - Cria amostras artificiais. C - Combina poda de dados com criação de amostras artificiais. D - Traduz os dados para outro idioma e depois de volta para o idioma original.

3.2.3 Modelo Computacional de Análise de Sentimentos

Como já mencionado, a análise de sentimentos é classificada como um problema de classificação, em que o objetivo é categorizar as opiniões expressas em textos em classes predefinidas, como por exemplo em positivo, negativo ou neutro (PANG; LEE, 2008; LIU, 2015). Nesse contexto, diversos algoritmos de classificação podem ser utilizados para construir modelos, como Máquinas de Vetores de Suporte (SVM), Regressão Logística, Redes Neurais, Árvores de Decisão, entre outros. A escolha do algoritmo depende de fatores como a natureza dos dados, a complexidade do problema e a necessidade de interpretar os resultados. Cada algoritmo apresenta suas vantagens e desvantagens, sendo necessário realizar uma análise criteriosa para selecionar o mais adequado para a tarefa específica de análise de sentimentos.

3.2.3.1 Classificadores

3.2.3.1.1 Regressão Logística

Apesar do nome, a regressão logística é um modelo linear amplamente utilizado para tarefas de classificação, e não de regressão. Esse modelo, descrito também no trabalho de NOGUEIRA (2021), aplica técnicas estatísticas para prever os valores de uma variável categórica, geralmente binária, com base em uma ou mais variáveis independentes, que podem ser contínuas e/ou binárias (GONZALEZ, 2018). Similar à regressão linear, a regressão logística ajusta pesos aos dados de treinamento para encontrar a melhor curva que se adapta aos dados, em vez de uma reta, como ocorre na regressão linear. Esse modelo calcula uma razão de probabilidade para a variável alvo, que é posteriormente transformada em uma escala logarítmica. Isso permite classificar observações com base na aproximação do valor correspondente (WITTEN; FRANK, 2005).

A equação utilizada para realizar a classificação com o classificador de Regressão Logística é expressa da seguinte maneira (SOUZA, 2017):

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{3.1}$$

onde e é o número de Euler, x_0 é o valor de x no ponto médio da curva, L é o valor máximo da curva e k a declividade da curva.

3.2.3.1.2 *Multinomial Naive Bayes*

Os métodos *Naive Bayes* fazem parte de um conjunto de algoritmos de aprendizado supervisionado que se baseiam na aplicação do teorema de Bayes, com a suposição "ingênua" (*naive*) de que as características são independentes entre si, dadas as classes. Essa suposição de independência condicional simplifica significativamente o cálculo das probabilidades,

tornando o algoritmo eficiente, mesmo em grandes conjuntos de dados. O teorema de Bayes (PEDREGOSA et al., 2011), descreve a seguinte relação entre a variável de classe y e o vetor de características $x = (x_1, x_2, ..., x_n)$:

$$P(y|x_1, \cdots, x_n) = \frac{P(y)P(x_1, \cdots, x_n|y)}{P(x_1, \cdots, x_n)} \tag{3.2}$$

utilizando a condição de independência tem-se que:

$$P(x_i|y, x_i \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) = P(x_i|y) \tag{3.3}$$

para todo i, tem-se que a relação pode ser simplificada para:

$$P(y|x_1, \cdots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, \cdots, x_n)} \tag{3.4}$$

uma vez que $P(y|x_1, \cdots, x_n)$ é constante dada a entrada, pode-se utilizar a seguinte regra de classificação:

$$P(y|x_1, \cdots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \tag{3.5}$$

$$\hat{y} = \underset{y}{\texttt{arg max}} P(y)\prod_{i=1}^{n} P(x_i|y) \tag{3.6}$$

e dessa forma pode-se utilizar a estimativa de Máximo A Posteriori (MAP) para estimar $P(y)$ e $P(x_i|y)$ sendo o primeiro a frequência relativa da classe no conjunto de treinamento.

A variação *Multinomial Naive Bayes* do *Naive Bayes* é especialmente projetada para lidar com dados distribuídos de forma multinomial. Essa é uma das versões clássicas do algoritmo, amplamente aplicada na classificação de textos, devido à sua eficiência na modelagem de distribuições de contagem, como a frequência de palavras em documentos. A parametrização da distribuição é dada por vetores $\theta_y = (\theta_{y1}, \cdots, \theta_{yn})$ para cada classe y, sendo n o número de características e θ_{yi} a probabilidade $P(x_i|y)$ da característica i aparecer em uma determinada amostra pertencente a classe y.

Os parâmetros θ_{yi} são estimados por meio de uma versão suavizada da probabilidade máxima, utilizando a contagem de frequência relativa, dada pela seguinte expressão:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{3.7}$$

onde $N_{yi} = \sum_{x \in T} x_i$ é o número de vezes que a característica i aparece na amostra da classe y no conjunto de treinamento T, e $N_y = \sum_{i}^{n} N_{yi}$ é a contagem total de todas as características da classe y.

3.2.3.1.3 *OneVSOneClassifier* e *OneVSRestClassifier*

A estratégia chamada de um contra um (*OneVsOne*), consiste em ajustar um classificador por par de classes (PEDREGOSA et al., 2011). No momento da previsão, a classe que recebeu mais votos é selecionada como a saída final. Uma vez que necessita do ajuste de $\frac{n\_classes \cdot (n\_classes-1)}{2}$ classificadores, o que o torna mais lento do que *OneVsRest*, devido à sua complexidade **O** ($n\_classes^2$). Contudo, essa abordagem pode ser vantajosa para algoritmos baseados em *kernel*, que têm dificuldades em escalar com grandes quantidades de amostras. Isso ocorre porque, no *OneVsOne*, cada problema de aprendizagem envolve apenas um pequeno subconjunto de dados, enquanto no *OneVsRest*, o conjunto completo de dados é utilizado *n_classes* vezes.

Por outro lado, o classificador denominado um contra todos (*OneVsRest*), segundo PEDREGOSA et al. (2011), ajusta um classificador para cada classe, tratando cada classe como a classe positiva, enquanto as demais são consideradas negativas. Este método é mais eficiente e apresenta como vantagem a sua interpretabilidade, pois cada classe é representada por um único classificador. Dessa forma, é possível obter informações detalhadas sobre uma classe específica ao inspecionar seu classificador correspondente.

3.2.3.1.4 *KNN - K Nearest Neighbors*

A classificação baseada em vizinhos, conforme descrito por PEDREGOSA et al. (2011), é uma abordagem de aprendizado baseada em instâncias, também conhecida como aprendizado não generalizante. Esse método não busca construir um modelo interno generalizado; em vez disso, armazena diretamente as instâncias dos dados de treinamento. Aqui a classificação é realizada por meio de uma votação majoritária simples entre os vizinhos mais próximos de um ponto: o ponto de consulta é classificado na classe que possui o maior número de representantes entre seus vizinhos mais próximos.

Dois classificadores distintos são oferecidos pelo *scikit-learn* (i) o *KNeighborsClassifier*, que realiza o aprendizado considerando os k-vizinhos mais próximos de cada ponto de consulta, sendo k um valor inteiro definido pelo usuário e (ii) o *RadiusNeighborsClassifier*, que utiliza o número de vizinhos localizados dentro de um raio fixo r em torno de cada ponto de treinamento, onde r é um valor em ponto flutuante especificado pelo usuário.

3.2.3.1.5 *Random Forest*

O *Random Forest Classifier* é um modelo de aprendizado supervisionado baseado em árvores de decisão que utiliza o método de *ensemble* para melhorar a precisão e robustez da classificação (PEDREGOSA et al., 2011). A ideia central do *Random Forest* é a de combinar múltiplas árvores de decisão independentes, cada uma treinada com subconjuntos aleatórios dos dados e das características.

Essa abordagem, conhecida como "*bagging*" (*bootstrap aggregating*), ajuda a reduzir o risco de *overfitting* e melhora a capacidade do modelo de generalizar para novos dados. Cada árvore no modelo realiza uma votação para determinar uma classe, e a classificação final é definida pela maioria dos votos entre as árvores.

3.2.3.1.6 *Linear SVC - Linear Support Vector Classification*

O *Linear Support Vector Classification* (*Linear SVC*) é uma implementação de classificadores de máquinas de vetores de suporte (SVM) que utiliza um *kernel* linear para separar dados em diferentes classes. Ele busca encontrar o hiperplano que maximiza a margem entre as classes no espaço de características, garantindo uma classificação eficiente e robusta, especialmente em problemas linearmente separáveis (PEDREGOSA et al., 2011). O *Linear SVC* é uma abordagem eficiente para grandes conjuntos de dados e oferece suporte a regularizações L1 e L2, permitindo flexibilidade no ajuste do modelo.

3.2.3.1.7 *SVM - Support Vector Machine*

O *Support Vector Machine* (SVM) é uma técnica de aprendizado supervisionado que pode ser usada tanto para classificação quanto para regressão. O SVM utiliza o conceito de vetores de suporte para encontrar o hiperplano ótimo que separa as classes no espaço de características (PEDREGOSA et al., 2011). Diferentemente do *Linear SVC*, a SVM suporta diversos tipos de *kernels*, como *radial basis function* (RBF), polinomial e *sigmoid*, permitindo lidar com problemas mais complexos, onde as classes não são linearmente separáveis. O SVM é amplamente utilizado devido à sua eficácia em dados de alta dimensionalidade e capacidade de generalização.

3.2.3.1.8 *Ensemble Voting Classifier*

Os métodos de *ensemble* (em tradução livre do inglês, "conjunto") consistem na combinação de diferentes classificadores de aprendizado de máquina, com o objetivo de aprimorar a precisão das previsões. Essa abordagem utiliza estratégias como voto majoritário ou cálculo da média das probabilidades previstas para determinar os rótulos das classes (PEDREGOSA et al., 2011). Por integrar modelos conceitualmente distintos, o *ensemble* demonstra-se especialmente eficaz para compensar as limitações individuais de cada classificador, resultando em um desempenho global superior.

3.3 *Business Intelligence*

O conceito de *Business Intelligence* (BI), ou inteligência de negócios, tem ganhado destaque na atualidade como uma ferramenta essencial para a gestão estratégica de organizações. De acordo com DUAN; XU (2012), o BI é definido como o processo de

transformar dados brutos em informações valiosas, capazes de gerar maior efetividade estratégica, *insights* operacionais e benefícios tangíveis para o processo de tomada de decisão nos negócios. Essa abordagem permite que as empresas identifiquem padrões, tendências e oportunidades, otimizando suas operações e aumentando sua competitividade no mercado.

NETO; CAMPOS (2023) destacam que as tecnologias de BI simplificam os processos de coleta, análise e entrega de informações, conhecidos como *Extraction, Transformation, Loading* (ETL). Após a transformação e limpeza dos dados, as informações são disponibilizadas de forma estruturada, prontas para subsidiar a tomada de decisões em temas ou assuntos relacionados ao negócio.

Com o avanço da era dos dados, as técnicas de BI têm sido aplicadas em diversos contextos para otimizar processos e embasar tomadas de decisão estratégicas. OLIVEIRA et al. (2023) implementaram BI na gestão da cadeia de suprimentos de uma empresa, reduzindo em 54,95% o tempo de execução da lista crítica, uma atividade que não agregava valor ao negócio. Já LINKE (2024) analisou os benefícios práticos do BI na gestão hospitalar, destacando a estruturação e visualização de dados, além do acesso em tempo real a informações críticas, o que facilita decisões gerenciais estratégicas. Em outro estudo, RANGEL (2024) desenvolveu uma plataforma de BI para analisar o perfil estudantil nos cursos superiores do IFPE – Campus Igarassu, permitindo identificar padrões relacionados à renda, gênero e taxas de evasão, contribuindo para um planejamento mais eficiente e ações de mitigação desses eventos.

No contexto do mercado consumidor, o BI também desempenha um papel fundamental. Como destacado por BRANDÃO (2016), compreender de forma clara o segmento de atuação de uma empresa é essencial para identificar quem são seus consumidores e como eles se comportam. Duas abordagens de segmentação se destacam como ferramentas essenciais para a análise de mercado: a segmentação demográfica, que categoriza o público com base em características populacionais como sexo, idade e renda, e a segmentação comportamental, que aprofunda a análise ao considerar conhecimento, atitudes, preferências e padrões de consumo (BRANDÃO, 2016).

No estudo de NOGUEIRA et al. (2024) sobre o consumo de queijos no Brasil, os autores exploraram questões relacionadas às duas abordagens de segmentação. Na segmentação demográfica, foi possível responder a perguntas como: (i) quem consome queijo? e (ii) em quais regiões o consumo é mais expressivo?. Já na segmentação comportamental, que investiga aspectos mais específicos do comportamento do consumidor, foram analisadas questões como: (i) quando ocorre esse consumo?, (ii) qual é o sentimento associado ao produto?, (iii) quais tipos de queijos se destacam? e (iv) quais são os principais acompanhamentos?.

Essas questões fornecem insumos importantes para que empresas e pesquisadores

compreendam melhor as preferências dos consumidores em um determinado segmento, possibilitando a identificação de tendências emergentes e a adaptação estratégica a um mercado em constante evolução. Dessa forma, o BI consolida-se como uma ferramenta essencial para aumentar a competitividade e a relevância no setor, ao integrar dados demográficos e comportamentais, identificar oportunidades e aprimorar a capacidade de resposta às demandas do mercado.

4 MATERIAL E MÉTODOS

Este capítulo apresenta de forma detalhada as etapas, ferramentas e abordagens utilizadas para a execução deste trabalho. Neste capítulo, são descritos os procedimentos adotados para coleta, processamento e análise dos dados, bem como as tecnologias e metodologias empregadas para garantir a qualidade, a robustez e a visualização dos resultados obtidos. A sistematização aqui apresentada busca oferecer uma visão clara e fundamentada do processo científico que embasa este estudo, assegurando a sua reprodutibilidade e transparência.

4.1 Arquitetura Computacional e Fluxo de Processamento do OC

A arquitetura do OC é estruturada em cinco módulos principais: (i) coleta de dados, (ii) armazenamento de dados, (iii) processamento e extração de informações a partir de dados de imagem, (iv) processamento e extração de informações de dados textuais e (v) pós-processamento e visualização dos resultados. A Figura 12 ilustra o fluxo completo de tarefas executadas pelo OC, que serão explicadas em maiores detalhes nas próximas subseções.

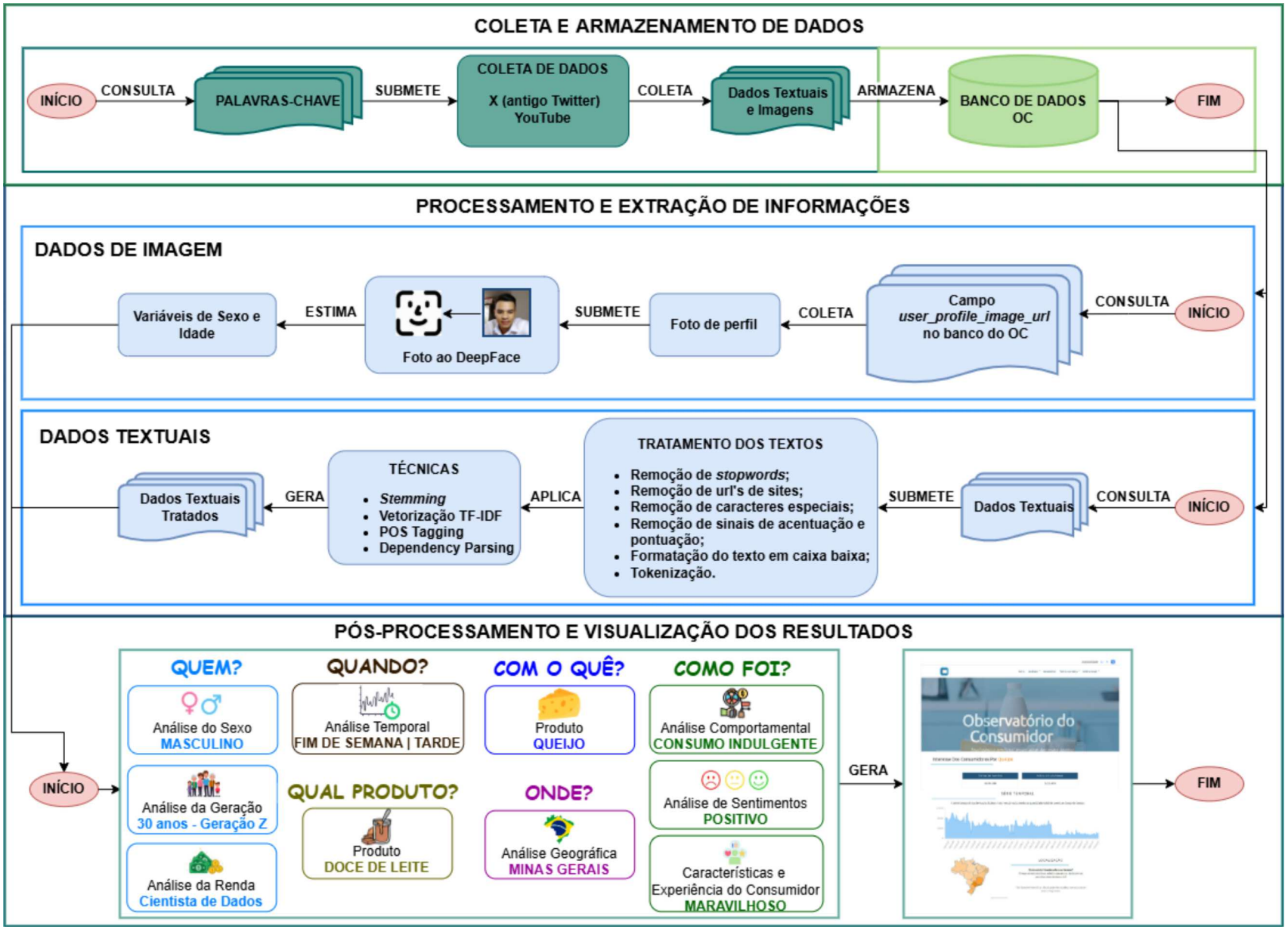


Figura 12 – Representação esquemática do fluxo de tarefas realizado neste trabalho. **Fonte:** autor.

4.1.1 Coleta de dados

O desenvolvimento do sistema do OC passou por várias fases até atingir sua versão atual. No final de 2019, o projeto foi lançado como um Mínimo Produto Viável (MPV), com foco na análise de dados do X sobre queijos artesanais brasileiros. O objetivo dessa fase inicial foi o de validar a viabilidade do projeto, identificando as principais informações a serem extraídas e os desafios a serem superados para expandir a análise a outras categorias de produtos lácteos.

Na versão inicial NOGUEIRA (2021) destacou que o sistema operava com um conjunto restrito de dados com apenas 82.959 postagens. Essa limitação estava relacionada à especificidade das palavras-chave utilizadas na coleta, o que reduziu a abrangência da análise. Assim, a ferramenta na primeira versão apresentava três resultados principais: (i) análise de sentimentos das publicações, (ii) características dos queijos artesanais e (iii) hábitos de consumo predominantes.

Com a validação da primeira versão, que utilizava 179 palavras-chave no contexto dos queijos artesanais, a expansão do sistema foi impulsionada para abranger outras categorias de produtos lácteos, tornando-se este o foco principal deste trabalho. Assim, foram adicionadas 114 novas palavras-chave, totalizando 293 ao todo, que foram organizadas em dez categorias, Bebidas Lácteas (5 palavras-chave), Creme de Leite (3), Doce de Leite (2), Iogurte (5), Leite (15), Leite Condensado (3), Leite Fermentado (5), Manteiga (6), Queijos (239) e Sorvete (10). A listagem completa de palavras-chave pode ser consultada no repositório do *GitHub*[1].

Essa adição ampliou significativamente o volume de dados coletados, armazenados e processados. Vale ressaltar que semelhante à primeira versão do OC, essa listagem foi construída com a supervisão e orientação de especialistas da Embrapa. Além dos dados provenientes do X, duas novas fontes de dados foram incorporadas ao sistema: Google Trends e YouTube. Essa expansão exigiu uma reestruturação abrangente nos processos de coleta, armazenamento e exibição de dados, com o objetivo de proporcionar maior eficiência e agilidade no acesso às informações pelos usuários finais do OC. Deste modo, para a rede social X, foram utilizadas 293 palavras-chave, enquanto que para o YouTube e Google Trends, foram empregadas apenas 10 palavras-chave, correspondentes às 10 categorias de lácteos de interesse. Essa diferença decorre das limitações das APIs de coleta de dados dessas plataformas.

Para realizar a coleta dos dados do X, um *script* na linguagem de programação R (**R** CORE TEAM, 2020) foi desenvolvido. Esse código foi responsável por importar a biblioteca *rtweet* (MW, 2019) realizar a autenticação de desenvolvedor na plataforma e submeter cada uma das palavra-chaves de interesse na API da mídia social para a realização do *download* dos dados. Como descrito no trabalho de NOGUEIRA (2021) a

[1] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/palavras_chave.md

API fornecida pelo X possuía limitações relacionadas com períodos de coleta de dados. A conta gratuita permitia a coleta de dados de até 7 dias anteriores. Dessa forma, para construir um conjunto de dados históricos, semanalmente, este *script* foi executado.

Para realizar a coleta de dados no YouTube, foi desenvolvido um *script* em Python utilizando as bibliotecas PyTube (PYTUBE, 2015) e a API oficial do Google denominada YouTube Data API v3[2] (GOOGLE, 2013). O processo tem início com o carregamento de palavras-chave previamente armazenadas no banco de dados, que são enviadas ao *script* para buscar informações iniciais sobre vídeos relacionados, utilizando a biblioteca PyTube. Essa etapa é fundamental para identificar os vídeos de interesse e obter seus identificadores únicos (*id_video*), os quais são posteriormente utilizados para recuperar informações mais detalhadas e complementares por meio da YouTube Data API v3.

Com o PyTube, são coletadas informações preliminares sobre os vídeos, como: *id_video*, *id_video_url*, título, URL do vídeo, duração em segundos, quantidade de visualizações, data de publicação e autor. Posteriormente, a YouTube Data API v3 é utilizada para complementar os dados. Nesse processo, o *id_video* é extraído da URL do vídeo e utilizado para acessar informações adicionais, como comentários (limitados a 30 páginas devido às restrições da conta gratuita da API oficial, que permite até 1000 solicitações), número de curtidas, número de comentários, descrição, categoria do vídeo e outras características que não são acessíveis diretamente pelo PyTube, como detalhes mais aprofundados sobre os comentários.

No caso do Google Trends, a coleta de dados é realizada por meio de *widgets*[3], que requerem a configuração prévia das consultas a serem disponibilizadas para os usuários finais do OC. Esse processo permite que os usuários realizem suas próprias buscas no Google Trends diretamente no sistema do OC, sem a necessidade de sair da plataforma. Isso garante maior flexibilidade na exploração dos dados, além de melhorar a interação com o sistema e possibilitar uma integração dinâmica das informações coletadas.

4.1.2 Armazenamento dos dados

Uma das diferenças mais marcantes da primeira para a segunda versão do OC está na forma de armazenar os dados, tanto os coletados quanto os processados pelas análises implementadas, para posterior exibição no sistema *web*. A primeira versão do OC a quantidade de dados coletados foi bastante reduzida. NOGUEIRA (2021) menciona que, em média, na primeira versão apenas 1.997 *tweets* eram coletados a cada semana.

[2] *Link* da YouTube Data API v3 https://developers.google.com/youtube/v3. Acesso em: 13 fev. 2025

[3] Um *widget* é um componente de interface gráfica que permite interação do usuário com um sistema ou aplicação (DIGITAL APP, 2025). Ele pode incluir elementos como botões, menus, barras de rolagem, caixas de texto, gráficos, entre outros, facilitando a execução de tarefas específicas e a navegação em interfaces digitais.

Com a adição de outras categorias de derivados lácteos, esse número cresceu bastante. Em média semanalmente 134.859 foram coletados, fazendo com que adaptações e construção de novas tabelas fossem realizadas visando melhorar o tempo de resposta às requisições realizadas e a experiência dos usuários do OC.

A versão atual do OC abrange uma ampla gama de produtos de interesse. Para atender a essa diversidade, as tabelas do banco de dados foram projetadas para serem genéricas. Essa modelagem flexível possibilita a análise de qualquer tipo de produto, inclusive fora do contexto lácteo, desde que as adaptações necessárias sejam devidamente implementadas.

4.1.2.1 Diagrama de entidade e relacionamento da aplicação

Após a identificação dos dados a serem inseridos e armazenados no Sistema de Gerenciamento de Banco de Dados (SGBD) do OC, foi desenvolvida uma nova modelagem, baseada na proposta inicial apresentada por NOGUEIRA (2021). Como a nova versão abrange dados além dos provenientes do X, incluindo também os do YouTube, foi necessário ajustar e criar modelagens distintas para os bancos de dados, com o objetivo de contemplar tanto as redes sociais X quanto o YouTube. As Figuras 13 e 14, juntamente com o repositório no *GitHub*[4], apresentam o relacionamento entre os dados e detalham a estrutura dos atributos das tabelas de cada plataforma.

[4] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/tabelas_armazenamento_dados.md

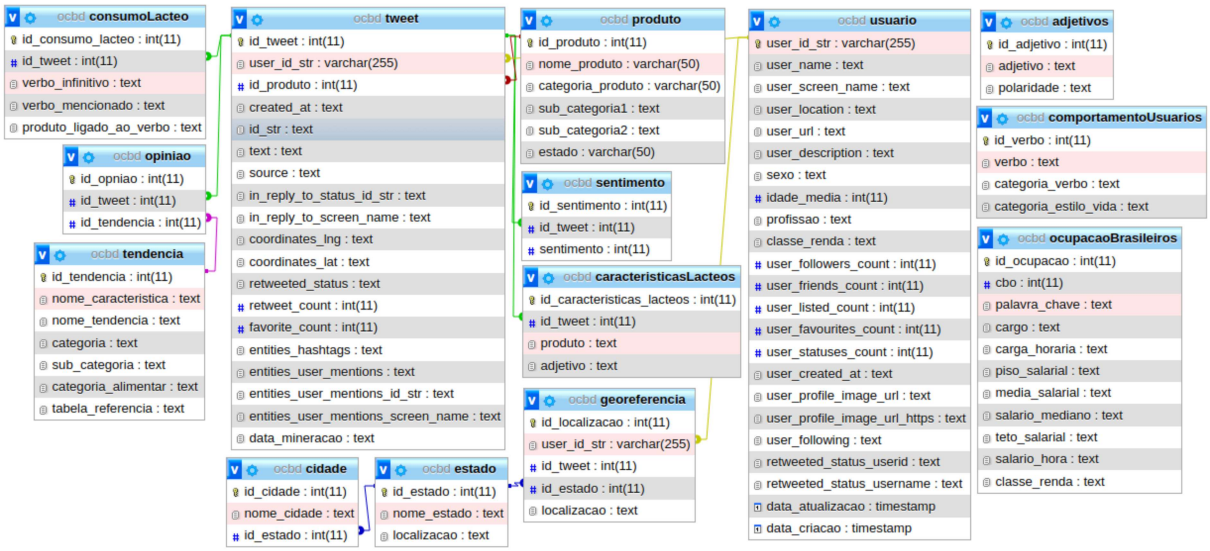


Figura 13 – Modelo de entidade e relacionamento adaptado do trabalho de NOGUEIRA (2021) referente à nova versão do banco de dados relacional da plataforma do "Observatório do Consumidor" para a rede social X. **Fonte:** autor.

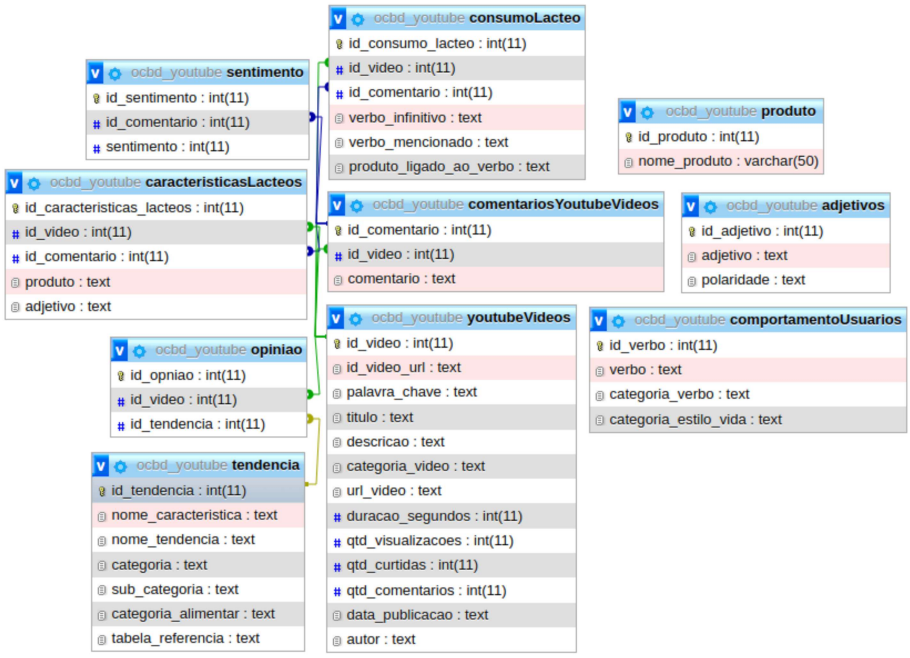


Figura 14 – Modelo de entidade e relacionamento adaptado do trabalho de NOGUEIRA (2021) referente à nova versão do banco de dados relacional da plataforma do "Observatório do Consumidor" para a rede social YouTube. **Fonte:** autor.

Para a rede social X, os dados das postagens são inseridos na tabela "*tweet*", enquanto as informações do usuário que realizou a postagem são armazenadas na tabela "usuario". A tabela "youtubeVideos", por sua vez, é responsável por armazenar as informações relacionadas aos vídeos postados no YouTube. Assim como na versão anterior do OC, a tabela "opiniao" foi mantida para armazenar, os resultados obtidos após o processamento de extração de informações, das postagens ou comentários do YouTube, armazenando características e hábitos de consumo de derivados lácteos. Além disso, as tabelas "georeferencia", "estado" e "cidade" foram criadas para otimizar o processo de identificação do estado e cidade de um usuário do X, com base na informação "*user_location*"[5] presente na tabela "usuario".

A tabela "produto" armazena os nomes dos produtos a serem consultados durante a coleta de dados. A tabela "tendencia", por sua vez, contém uma lista de 1.485 alimentos extraídos da Tabela Brasileira de Composição de Alimentos (TACO) (NEPA, 2011) e da Pesquisa de Orçamentos Familiares (POF) (IBGE, 2011). Esses alimentos foram classificados conforme o Guia Alimentar para a População Brasileira (MINISTÉRIO DA SAÚDE, 2014) em quatro categorias: alimentos *in natura* ou minimamente processados, alimentos processados, alimentos ultraprocessados e ingredientes culinários.

Em ambos os modelos de entidade-relacionamento, as tabelas "comportamentoUsuarios", "ocupacaoBrasileiros" e "adjetivos" funcionam como repositórios de consulta, armazenando listas de palavras-chave relevantes para a extração de informações específicas nos dados. Essas palavras-chave são utilizadas em processos de análise textual, auxiliando na obtenção de percepções valiosas. Uma vez populadas, essas tabelas são dedicadas exclusivamente a operações de consulta, fornecendo informações essenciais para as etapas subsequentes do processamento. A construção dessas tabelas foi planejada para otimizar tanto o armazenamento quanto o desempenho das operações, permitindo maior eficiência no manuseio dos dados.

Para a análise de comportamentos, foram criadas as tabelas "comportamentoUsuarios" e "consumoLacteo". A tabela "comportamentoUsuarios" armazena verbos classificados em seis categorias de ações: alimentação, gerenciamento do estresse, controle de tóxicos, atividades físicas, qualidade do sono e formação e manutenção de relacionamentos. Essas categorias são baseadas em recomendações da medicina do estilo de vida para a promoção de hábitos saudáveis. Já a tabela "consumoLacteo" armazena os resultados desse processamento, exibindo informações sobre o consumo de lácteos. Por fim, a tabela "sentimento" é responsável por armazenar a polaridade dos sentimentos presentes nas postagens ou comentários dos vídeos.

A integração com o Google Trends não demandou a criação de um banco de

[5] Informação fornecida pelo próprio usuário do X. Por ser um campo "aberto", que permite ao usuário digitar qualquer dado, requer um pré-processamento para identificação e validação da localização.

dados dedicado. Em vez disso, foi desenvolvida uma página que incorpora os *widgets* disponibilizados pela plataforma do Google. Essa página oferece aos usuários a possibilidade de selecionar até cinco produtos lácteos para comparação simultânea, respeitando a limitação imposta pelo próprio Google Trends.

A estrutura do banco de dados para o YouTube assemelha-se à do X, com a diferença de que a plataforma não permite a coleta de informações de localização dos autores de vídeos ou comentários. Consequentemente, as tabelas "cidade" e "estado" não são necessárias nesse caso. A tabela "produto", por sua vez, como já mencionado, contém apenas dez palavras-chave, correspondentes às dez categorias de produtos lácteos de interesse. Essa redução se deve à limitação da YouTube Data API v3: o uso das 293 palavras-chave, como na plataforma X, excederia as cotas de acesso aos dados, inviabilizando a coleta.

Um dos objetivos do OC é permitir que seus usuários finais estudem dados históricos coletados. Devido à grande quantidade de dados processados todas as semanas, como mais de 23 milhões de publicações na rede social X, o processamento em tempo real se mostrou inviável, resultando em tempos de resposta inadequados para os usuários. Para melhorar a experiência e proporcionar consultas mais eficientes e em tempo real, todos os resultados das análises realizadas pelo OC são processados *offline*, devido ao alto custo computacional, e armazenados nas tabelas de resultados semanais e mensais. Com isso, os dados do X e do YouTube foram organizados em tabelas com resultados pré-processados, garantindo maior agilidade nas consultas, pois as análises já são calculadas e armazenadas previamente. As Figuras 15 e 16 ilustram os modelos entidade-relacionamento dessas tabelas para as redes sociais X e YouTube, respectivamente.

As tabelas "identificacaoSemana" e "identificacaoMes" são pilares dessa estrutura. A primeira armazena informações semanais, como índices inicial e final de cada semana, total de publicações e data da coleta inicial dos dados do X. A segunda, por sua vez, registra o mês (número de 1 a 12), o ano do vídeo e o total de vídeos por mês para o YouTube.

Existem duas diferenças principais entre as tabelas de resultados do X e do YouTube. Primeiramente, a granularidade temporal: os dados do X são armazenados semanalmente, enquanto os do YouTube são mensais. Essa diferença se justifica pelo menor volume de dados do YouTube em comparação ao X, tornando a agregação mensal mais adequada para a visualização gráfica. Em segundo lugar, o processo de filtragem difere entre as plataformas. Devido ao grande volume de dados do X, a filtragem é possível em três dimensões: uma baseada no agrupamento de informações pela possível localização dos usuários, outra pelo sentimento expresso e outra por cada categoria de produto lácteo.

Para armazenar os resultados semanais filtrados por categoria de derivados lácteos e região, foram criadas 10 tabelas para a rede social X, cada uma representando uma categoria específica de produto, e 6 tabelas filtradas por região do Brasil, além de 3 tabelas

Figura 15 – Diagrama de entidade e relacionamento das tabelas de resultado da rede social X. **Fonte:** autor.

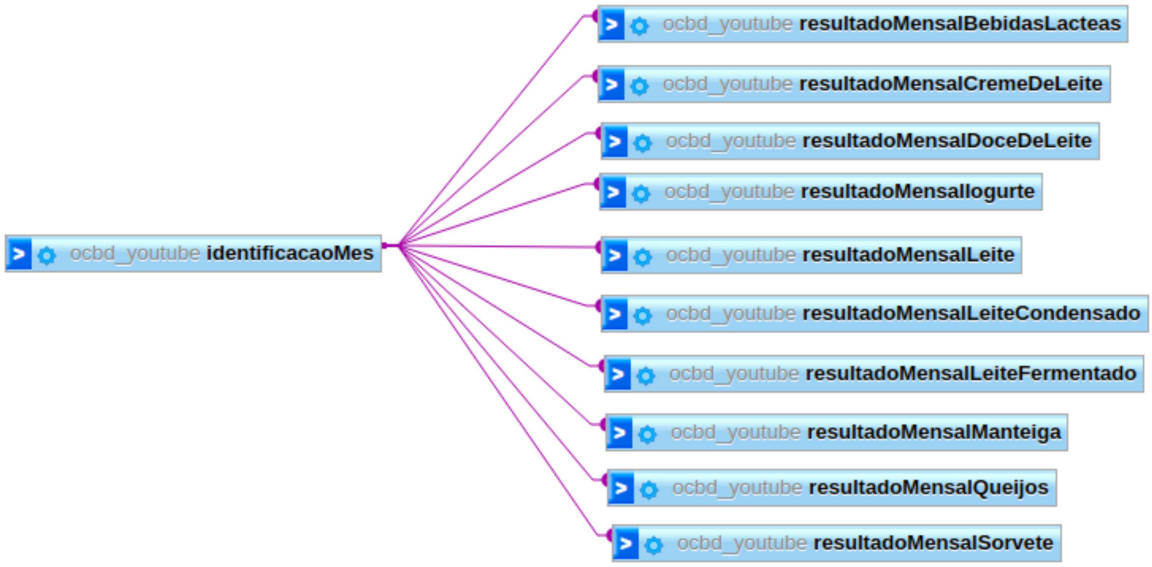


Figura 16 – Diagrama de entidade e relacionamento das tabelas de resultado da rede social YouTube. **Fonte:** autor.

para resultados filtrados por sentimento. Para a rede social YouTube, 10 tabelas mensais foram criadas, agrupadas por categoria de derivado lácteo. Os atributos de cada tabela podem ser visualizados nos repositórios do *GitHub* com informações das tabelas do X[6] e do YouTube[7].

4.1.2.2 Diagrama de caso de uso da aplicação

Na Figura 17, está apresentado o diagrama de caso de uso da aplicação, que ilustra a interação dos usuários com as principais funcionalidades oferecidas pela ferramenta para os dois atores que a utilizam.

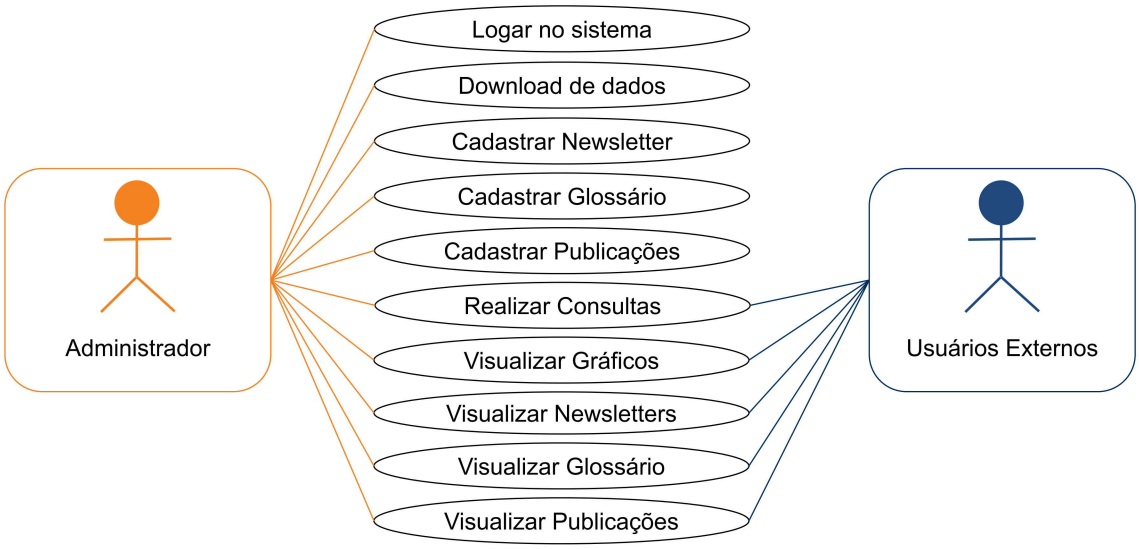


Figura 17 – Diagrama de caso de uso da aplicação. **Fonte:** autor.

No diagrama de caso de uso Figura 17, são identificados dois atores principais no sistema: o "Administrador" e o "Usuário Externo". O Administrador tem a responsabilidade de gerenciar todas as funcionalidades da plataforma, incluindo a administração de cadastros para *newsletters* mensais, a manutenção do glossário de termos técnicos, o registro de publicações científicas e o *download* das tabelas de resultados. Além dessas funções, o Administrador também pode executar todas as operações permitidas ao Usuário Externo, como realizar consultas na ferramenta, visualizar gráficos, acessar *newsletters* publicadas, consultar o glossário e referências de publicações.

Por outro lado, os "Usuários Externos" possuem permissões mais restritas, focadas apenas na consulta e visualização de informações. Eles podem utilizar a ferramenta para

[6] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/tabelas_resultados_semanais.md

[7] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/tabelas_resultados_mensais.md

visualizar gráficos, acessar *newsletters* prontas, consultar o glossário e acessar publicações científicas. Essa hierarquia de permissões garante que os Administradores tenham controle total sobre a gestão e atualização do conteúdo do sistema, enquanto os Usuários Externos desfrutam de um acesso simplificado e organizado às informações necessárias.

4.1.3 Processamento e extração de informações a partir de dados de imagem

Com os dados devidamente armazenados, a etapa seguinte envolve a extração de informações relevantes a partir desses conjuntos. Para isso, o OC realiza uma série de processamentos e análises, abrangendo tanto dados de imagem quanto textuais, sendo estes últimos a principal fonte de informação neste trabalho.

4.1.3.1 Análise de sexo e geração em fotos de perfis

O OC implementa duas análises específicas para dados de imagem: a identificação de sexo e a estimativa da faixa etária dos usuários da rede social X. Idade e gênero são fatores sociodemográficos essenciais na análise do perfil do consumidor. No entanto, essas informações frequentemente não estão disponíveis nos dados fornecidos pelas APIs das redes sociais. Para suprir essa lacuna, o OC utiliza as imagens de perfil dos usuários do X que publicaram conteúdos relacionados a derivados lácteos para identificar essas características.

Para estimar as informações de idade e sexo dos usuários, foi desenvolvido um *script* que consultava o campo "*user_profile_image_url*" na tabela de usuários e tentava baixar as imagens de perfil em resolução de 400×400 *pixels*, caso estivessem disponíveis. As imagens obtidas eram então processadas pelo *framework DeepFace*[8] (SERENGIL; OZPINAR, 2021), uma ferramenta avançada em Python para reconhecimento facial e análise de atributos faciais, incluindo idade, sexo, emoção e raça. Esse *framework* utiliza uma combinação de modelos como *VGG-Face*, *Google FaceNet*, *OpenFace*, *Facebook DeepFace*, *DeepID*, *ArcFace*, *Dlib* e *SFace*, para gerar previsões sobre o sexo (feminino ou masculino) e a faixa etária dos usuários. As idades identificadas foram agrupadas em diferentes gerações, classificando os usuários da seguinte forma: Geração Silenciosa para os nascidos antes de 1945, Geração *Baby Boomers* para os nascidos entre 1946 e 1960, Geração X para os nascidos entre 1961 e 1980, Geração Y para os nascidos entre 1981 e 1995, Geração Z para os nascidos entre 1995 e 2010 e Geração *Alpha* para os nascidos após 2011.

Para o modelo de idade SERENGIL; OZPINAR (2021) menciona que para esta tarefa de previsão este modelo obteve um erro de $\pm 4,65$ para a métrica MAE, ou seja, o modelo estima as idades com mais ou menos 4,5 anos, o que é aceitável. Já para o modelo de previsão de sexo os resultados foram de 97,44% para a métrica *accuracy*, 96,29% para *precision* e 95,05% para *recall* que são resultados satisfatórios.

[8] https://github.com/serengil/deepface. Acesso em: 13 fev. 2025.

Além das estimativas de sexo e idade, o *framework DeepFace* também fornece informações sobre raça e emoção. No entanto, o OC optou por limitar suas análises apenas às características de idade e gênero. A complexa diversidade racial do Brasil, resultado da miscigenação, pode não ser adequadamente representada pela ferramenta, o que poderia levar a resultados pouco representativos. Além disso, a análise de emoções, baseada em fotos de perfil, foi considerada irrelevante para o projeto, uma vez que a expressão facial em uma foto de perfil não oferece dados significativos para a ferramenta.

4.1.4 Processamento e extração de informações a partir de dados textuais

Como mencionado anteriormente, os dados textuais foram os mais explorados pelo OC devido à sua predominância nos dados coletados. Para esse tipo de dado, foram realizadas análises de renda, temporal, de produtos lácteos ou não, geográfica, comportamental e de sentimentos. Para cada semana de coleta, é realizada uma consulta no banco de dados para recuperar as publicações relevantes. Cada postagem passa por etapas de pré-processamento que incluem tokenização, remoção de *stopwords* e normalização dos textos, convertendo-os para letras minúsculas (*lowercase*).

Além disso, foram aplicadas técnicas como *stemming*, que reduz as palavras aos seus radicais, e vetorização dos textos utilizando a abordagem *TF-IDF*, que identifica as palavras mais relevantes em cada postagem. Outras técnicas avançadas de processamento de linguagem natural, como *Part-of-Speech Tagging* (*POS Tagging*) e *Dependency Parsing*, também foram empregadas para analisar as estruturas gramaticais das postagens, identificando relações sintáticas e categorias gramaticais nas sentenças.

Esses procedimentos foram selecionados com base na revisão da literatura apresentado no capítulo 2, onde foram destacados como os mais eficazes e amplamente utilizados para o tratamento e análise de dados textuais, garantindo maior precisão e profundidade nas análises realizadas.

4.1.4.1 Análise de renda

Outro ponto importante para o perfil do consumidor é a renda. Para extrair as possíveis profissões dos usuários, aplicamos técnicas de mineração de dados no campo "*user_description*" da rede social X, a única plataforma deste estudo que oferece um espaço para autodescrição. Esse campo pode conter informações sobre as ocupações mencionadas pelos usuários, o que nos permite estimar sua renda e realizar uma análise mais aprofundada de seu perfil como consumidores.

Com base no estudo de PREOţIUC-PIETRO et al. (2015), que utilizou o *Standard Occupational Classification* (SOC), uma taxonomia padronizada de empregos desenvolvida pelo *Office of National Statistics* (ONS) no Reino Unido para mapear usuários do X. Neste trabalho adotou-se uma abordagem semelhante, porém, ajustada à realidade brasileira com

a Tabela Cargos e Salários 2024[9] – Piso Salarial das Profissões. Dessa forma, o conteúdo textual desse campo foi processado para identificar as profissões declaradas, permitindo a posterior correlação dessas ocupações com os salários médios estimados, utilizando essa tabela como referência.

A análise do campo "*user_description*", embora indireta, permite estimar a renda dos usuários que mencionam profissões de interesse, contribuindo para a construção de um perfil mais completo do consumidor. Ao correlacionar as profissões identificadas com os salários médios, podemos inferir a classe econômica dos usuários da rede social que mencionaram produtos lácteos, enriquecendo a análise do perfil do consumidor.

Para mantermos sempre cálculos atualizados acerca do salário mínimo vigente, um algoritmo é empregado para extrair informações atualizadas diretamente do site do Instituto de Pesquisa Econômica Aplicada (IPEA)[10]. Posteriormente, para categorizar as profissões dentro dos grupos de classes A, B, C, D e E, utilizamos a classificação do IBGE, conforme apresentada na tabela da Pesquisa de Orçamentos Familiares (IBGE, 2011). Essa classificação divide as classes da seguinte maneira: Classe E, até 2 salários; Classe D, de 2 a 3 salários; Classe C, de 3 a 6 salários; Classe B, de 6 a 10 salários; e Classe A, mais de 10 salários.

4.1.4.2 Análise temporal

A análise de dados históricos é fundamental para compreender eventos passados e embasar decisões futuras. Cada plataforma organiza suas informações de maneira distinta, permitindo diferentes níveis e abordagens de análise. Embora não seja focada diretamente no conteúdo textual, essa análise contabiliza a quantidade total de postagens coletadas no X e o número de vídeos no YouTube em que aparecem as palavras-chave de interesse.

No OC, os dados do X são apresentados em gráficos temporais organizados semanalmente, enquanto no YouTube a análise é realizada mensalmente. Já o Google Trends oferece flexibilidade com oito intervalos de tempo, variando da última hora aos últimos cinco anos. Essa diversidade na granularidade das análises entre as plataformas possibilita ao OC oferecer uma visão detalhada e abrangente sobre o comportamento dos consumidores em relação aos produtos lácteos ao longo do tempo.

4.1.4.3 Análise de produtos lácteos e demais produtos mencionados

Esta análise, conforme indicado no título, tem como objetivo implementar mecanismos para uma avaliação detalhada dos produtos lácteos, identificando as categorias que despertam maior interesse e os produtos frequentemente mencionados ou consumidos em conjunto com os lácteos. Para isso, o sistema do OC aplica uma série de processos de

[9] https://www.salario.com.br/tabela-salarial. Acesso em: 14 fev. 2025.
[10] http://www.ipeadata.gov.br/ExibeSerie.aspx?serid=1739471028. Acesso em: 14 fev. 2025.

análise textual nas fontes de dados estudadas, permitindo não apenas mapear padrões de consumo, mas também avaliar as combinações alimentares associadas.

A abordagem adotada possibilita classificar o consumo de lácteos com base nas combinações alimentares identificadas. Como os hábitos alimentares variam, o consumo de lácteos pode se enquadrar em padrões saudáveis ou não saudáveis, dependendo dos alimentos com os quais são combinados. Maiores detalhes sobre a implementação dessa análise serão apresentados na subseção 4.1.4.5, dedicada à "Análise Comportamental".

4.1.4.4 Análise geográfica

A localização é outro fator crucial na análise do perfil do consumidor. NOGUEIRA (2021) destaca que questões relacionadas ao regionalismo, como no caso dos queijos artesanais, fazem com que características e hábitos de consumo variem significativamente entre regiões, dificultando a definição de uma tendência geral de consumo para esse tipo de produto.

Entre as plataformas utilizadas no OC, apenas o Google Trends e o X oferecem mecanismos para extração de informações geográficas. O Google Trends facilita essa análise ao fornecer diretamente os estados brasileiros com maior interesse em determinados produtos lácteos. Por outro lado, no caso do X, a localização é preenchida livremente pelos usuários no campo *location*, exigindo um processamento adicional. Para identificar as cidades e estados mencionados, foi realizada uma etapa de mineração de dados nesse campo, utilizando a base de dados de cidades e estados brasileiros disponíveis no OC.

Para facilitar as análises regionais, as localizações extraídas do campo *location* foram processadas e convertidas para o formato padrão BR-sigla do estado brasileiro. Essa conversão foi necessária para tornar possível a exibição dos resultados na ferramenta computacional desenvolvida. Em seguida, essas informações foram agrupadas de acordo com as cinco grandes regiões do Brasil: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Essa abordagem possibilita uma visão mais estruturada e eficiente das diferenças regionais no consumo de produtos lácteos.

4.1.4.5 Análise comportamental

Outro aspecto que colabora para a análise do perfil do consumidor está relacionado com a identificação de estilos de vida, ou seja, seus comportamentos. Como descrito no trabalho de PHILLIPS et al. (2020), a teoria denominada "Medicina do Estilo de Vida" formalizada pelo Dr. RIPPE (1999) é definida como uma prática que se baseia em evidências que podem ajudar indivíduos e comunidades mudarem aspectos relacionados com seus estilos de vida (incluindo nutrição, atividade física, gerenciamento de estresse, apoio social e exposições ambientais) que visa prevenir, tratar e até reverter a progressão de doenças crônicas (SAGNER et al., 2014).

PHILLIPS et al. (2020) detalha em seu trabalho 6 principais comportamentos que estão relacionados com a medicina do estilo de vida, sendo alimentação saudável, atividades físicas, manejo ou controle do estresse, diminuição ou suspensão do uso de tabaco, melhores condições de sono e relacionamentos. Utilizando esta teoria para analisar comportamento dos usuários, do X e YouTube, que publicaram sobre os lácteos, uma listagem no repositório do *GitHub*[11] com verbos categorizados nas categorias Consumo de Alimentos, Sono, Atividade Física, Controle de Tóxicos, Relacionamentos e Controle do Estresse foi desenvolvida com intuito de mapear quais foram as ações mais realizadas por estes indivíduos.

Com a estrutura devidamente configurada, iniciou-se a extração de informações sobre os principais hábitos dos usuários. Para isso, foi utilizada a biblioteca *spaCy* (HONNIBAL; MONTANI, 2017), que possibilitou a implementação de análises de dependência sintática, fundamentais para compreender as relações gramaticais nos textos. O principal objetivo dessa abordagem foi identificar ações específicas realizadas pelos usuários, relacionadas a produtos lácteos, por meio da análise das conexões entre verbos de interesse e substantivos que representam esses produtos. Por exemplo, observou-se verbos que indicam ações de “comprar”, “consumir”, “evitar” ou “recomendar” muito associados a produtos como “queijo”, “leite” e “iogurte” que neste caso eram substantivos. Ao mapear essas relações verbo-substantivo, torna-se possível obter uma visão mais profunda sobre as atitudes dos usuários destes ambientes virtuais em relação a esses produtos.

4.1.4.5.1 Hábitos de consumo

Para identificar os principais hábitos de consumo dos derivados lácteos no Brasil, 17 categorias foram criadas com base em nomes de produtos oriundos de duas tabelas, (i) a POF (IBGE, 2011) e da (ii) TACO (NEPA, 2011). Esta listagem foi submetida a um processo de padronização para remoção de produtos duplicados. Os nomes dos produtos foram agrupados em: Bebidas alcoólicas, Bebidas não alcoólicas e infusões, Carnes e derivados, Cereais e derivados, Enlatados e conservas, Farinhas, féculas e massas, Frutas e derivados, Gorduras e óleos, Laticínios, Nozes e sementes, Outros alimentos preparados, Ovos e derivados, Panificados, Pescados e frutos-do-mar, Produtos açucarados, Sais e condimentos, Verduras, hortaliças e derivados.

Para caracterizar o consumo de lácteos como saudável ou não, considerando os diversos tipos de alimentos associados a cada categoria, foram utilizadas as definições estabelecidas no Guia Alimentar para a População Brasileira (GAPB) (MINISTÉRIO DA SAÚDE, 2014). Assim, os nomes dos produtos foram classificados em uma das quatro

[11] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/analise_comportamental.md. Acesso em: 14 fev. 2025.

categorias definidas pelo guia: (i) alimentos *in natura* ou minimamente processados, (ii) ingredientes culinários, (iii) alimentos processados e (iv) alimentos ultraprocessados.

Os alimentos classificados como *in natura* ou minimamente processados de acordo com o MINISTÉRIO DA SAÚDE (2014) são os recomendados para compor a base de alimentos a serem consumidos pelos brasileiros. Os alimentos *in natura* são definidos por MINISTÉRIO DA SAÚDE (2014) como aqueles que são obtidos diretamente de plantas ou de animais, não sofrendo nenhuma alteração após deixar a natureza. Já os alimentos minimamente processados, são definidos como os alimentos *in natura* que são submetidos a processos de limpeza, remoção de partes não comestíveis ou indesejáveis, fracionamento, moagem, secagem, fermentação, pasteurização, refrigeração, congelamento e processos similares que não envolvam agregação de sal, açúcar, óleos, gorduras ou outras substâncias ao alimento original.

Os alimentos classificados como ingredientes culinários, conforme o MINISTÉRIO DA SAÚDE (2014), são os produtos extraídos de alimentos *in natura* ou da natureza por meio de processos como a prensagem, a moagem, a trituração, a pulverização e o refino. É recomendado que a utilização destes produtos seja feita em pequenas quantidades. Em geral, são usados para temperar, cozinhar alimentos e criar as mais diversas e variadas receitas culinárias, incluindo os caldos e as sopas, as saladas, as tortas, os pães, os bolos, os doces e as conservas.

Os alimentos processados, segundo o MINISTÉRIO DA SAÚDE (2014), são fabricados pela indústria que adiciona sal, açúcar ou outra substância de uso culinário a alimentos *in natura* para torná-los duráveis e mais agradáveis ao paladar. São produtos derivados diretamente de alimentos, sendo reconhecidos como versões dos alimentos originais. São usualmente consumidos como parte ou acompanhamento de preparações culinárias feitas com base em alimentos minimamente processados.

Já os alimentos ultraprocessados, também de acordo com o MINISTÉRIO DA SAÚDE (2014) devem ser evitados, sendo produzidos por meio de formulações industriais feitas inteiramente ou majoritariamente de substâncias extraídas de alimentos (óleos, gorduras, açúcar, amido, proteínas), derivadas de constituintes de alimentos (gorduras hidrogenadas, amido modificado) ou sintetizadas em laboratório com base em matérias orgânicas como petróleo e carvão (corantes, aromatizantes, realçadores de sabor e vários tipos de aditivos usados para dotar os produtos de propriedades sensoriais atraentes). Técnicas de manufatura incluem extrusão, moldagem, e pré-processamento por fritura ou cozimento. Os alimentos pertencentes às quatro categorias mencionadas podem ser consultados em detalhes no repositório do *GitHub*[12].

[12] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/categoria_produtos_gapb.md. Acesso em: 14 fev. 2025.

4.1.4.5.2 Características dos produtos lácteos

Para identificar as características dos produtos lácteos, foram definidas cinco categorias adaptadas e criadas com base no trabalho de NOGUEIRA et al. (2021) e na análise da frequência das palavras mais mencionadas nas publicações. Os termos identificados foram organizados nas seguintes categorias: Embalagem, Informações do produto, Composição, Produção e Contaminação. A listagem completa dos principais termos referentes hábitos e características relacionadas aos produtos lácteos está disponível em detalhes no repositório do *GitHub*[13].

4.1.4.5.3 Adjetivos

No estudo de NOGUEIRA et al. (2021), foram definidas categorias como cor, formato, peso, tamanho, tipo, odor, preço, sabor e textura. Ao analisá-las em detalhe, observou-se que os termos atribuídos a essas categorias eram, na verdade, adjetivos utilizados para qualificar os produtos lácteos. Com isso, optou-se por reorganizar esses termos na categoria de adjetivos, tornando a análise mais objetiva e direcionada.

Foi desenvolvida uma listagem específica de adjetivos para identificar as qualidades positivas, negativas e neutras mais mencionadas pelos usuários do X e YouTube em relação aos produtos lácteos. Para isso, utilizou-se a biblioteca *spaCy* e uma amostra de 651.598 *tweets*, que passaram por um processo de *POS-Tagging*. Para essa etapa, foi empregado o modelo pré-treinado da língua portuguesa *pt_core_news_lg* da biblioteca *spaCy*, reconhecido por sua alta acurácia para identificação de adjetivos associados aos produtos lácteos.

Posteriormente, os adjetivos identificados foram submetidos à plataforma GOTIT, para que cada adjetivo fosse classificado com relação às polaridades como sendo negativas, neutra ou positiva. Após essa etapa automatizada, foi realizado um processo manual de validação para garantir a precisão das classificações, preparando os dados para uso na ferramenta *web* do OC, em que os adjetivos mais frequentes são representados em uma *word cloud* (nuvem de palavras) organizada por categoria. A listagem completa dos adjetivos mais citados e suas respectivas categorizações está disponível no repositório do *GitHub*[14].

4.1.4.6 Análise de sentimentos

Outra análise que o OC implementa é a de sentimentos. Essa análise desempenha um papel fundamental na formulação de estratégias de *marketing* e no desenvolvimento

[13] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/tendencia.md. Acesso em: 14 fev. 2025.

[14] https://github.com/Thallys-nogueira/TeseDoutorado/blob/main/lista_adjetivos.md. Acesso em: 14 fev. 2025.

de produtos, uma vez que permite a extração automatizada de opiniões, sentimentos e emoções expressas em texto (NARAYANAN et al., 2009). Antes dos anos 2000, LIU (2012) observou que a disponibilidade de fontes de textos de opinião em formato digital era limitada, o que dificultava esse tipo de análise. No entanto, ao longo dos anos, esse cenário passou por mudanças significativas. Conforme apontado por BRITO (2017), o advento da internet e a proliferação de documentos online, especialmente em mídias sociais, transformaram a análise de sentimentos em uma das áreas mais dinâmicas e promissoras do processamento de linguagem natural.

Entender a polaridade dos sentimentos expressos nos textos é fundamental para obter uma compreensão mais aprofundada das opiniões dos usuários e dos comportamentos que eles adotam. Saber se um texto expressa um sentimento positivo, negativo ou neutro nos permite avaliar, por exemplo, se uma determinada ação foi bem recebida ou não, e se o consumo de um produto foi satisfatório para o consumidor. Ao correlacionar os sentimentos expressos pelos consumidores com suas ações e comportamentos, as empresas podem estimar possíveis impactos de suas estratégias e ajustar suas abordagens em tempo real, otimizando a experiência do cliente e fortalecendo sua posição competitiva no mercado.

A análise de sentimentos foi aplicada aos dados coletados das redes sociais X e YouTube. No caso do X, o modelo foi utilizado para classificar o conteúdo presente no campo "text" das postagens, enquanto, no YouTube, a análise foi realizada sobre os comentários armazenados no campo "comentario" de vídeos selecionados. O modelo responsável por essa classificação foi desenvolvido por NOGUEIRA et al. (2024) com o propósito de aprimorar a detecção de sentimentos em publicações relacionadas a produtos lácteos. O processo completo de construção do modelo, incluindo seus detalhes metodológicos, será descrito nas próximas subseções. A Figura 18 apresenta uma visão geral do processo metodológico adotado.

4.1.4.6.1 Seleção e rotulagem automática de dados

Para criar o conjunto de treinamento do modelo, NOGUEIRA et al. (2024), utilizaram os dados coletados da rede social X, seguindo o processo apresentado na subseção 4.1.1. No total, foram selecionados 96.645 *tweets*, dos quais 86.060 foram destinados à composição dos conjuntos de treinamento e validação. Os 10.585 *tweets* restantes foram reservados para formar o conjunto de teste, com o objetivo de avaliar a capacidade de generalização do modelo em dados nunca antes vistos.

O processo metodológico descrito por PINTO; ROCIO (2019) foi adotado para a rotulagem automática das amostras nas categorias negativa, neutra e positiva. Para isso, foram empregadas quatro APIs de análise de sentimentos mencionadas pelo próprio autor: *Amazon Comprehend*, *Google Natural Language*, *IBM Watson Natural Language Understanding* e *Microsoft Text Analytics*. Além dessas, a API *GOTIT*, utilizada no estudo

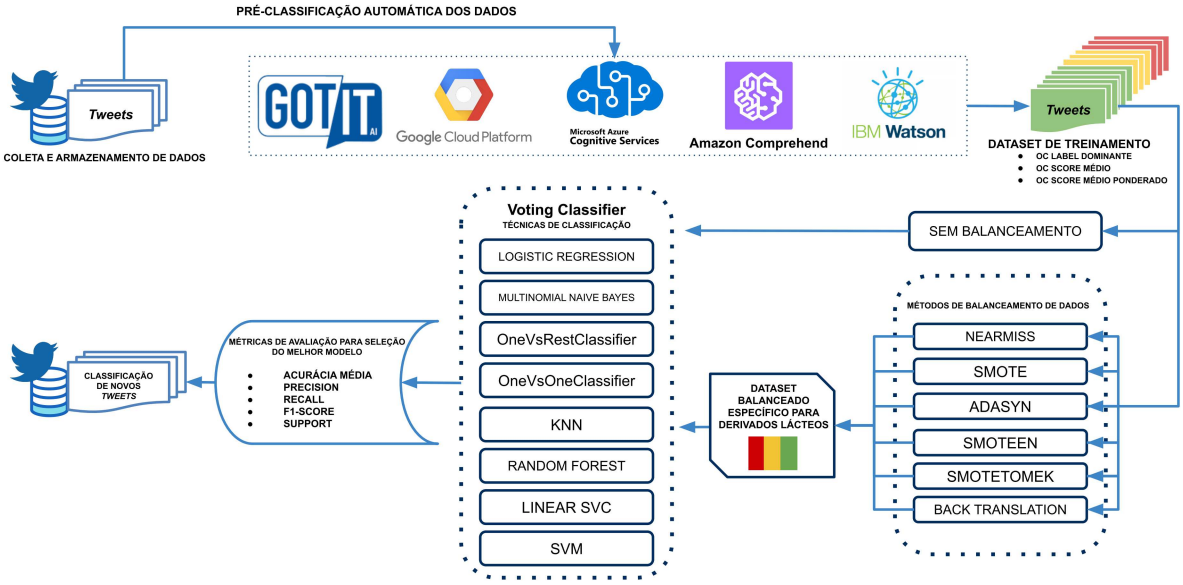


Figura 18 – Processo metodológico implementado para construção do modelo de análise de sentimentos no trabalho de NOGUEIRA et al. (2024). A fase inicial consistiu na coleta de *tweets* para a construção do conjunto de dados. Após serem coletados e devidamente armazenados, esses *tweets* foram submetidos à classificação por cinco APIs de processamento de linguagem natural, com o objetivo de categorizar automaticamente os dados (PINTO; ROCIO, 2019). Em seguida, os conjuntos de dados foram organizados com base em suas polaridades, aplicando-se seis técnicas de balanceamento de dados. É importante destacar que o método de *back translation* foi aplicado apenas ao conjunto de dados *OC Label Dominante*, devido ao seu alto custo computacional. Após a etapa de balanceamento de dados, foram desenvolvidos modelos de análise de sentimentos específicos para cada conjunto de dados. Para avaliar esses modelos, utilizou-se a estratégia de *Voting classifier*, e o desempenho foi medido por meio de quatro métricas. Esse processo teve como objetivo identificar o modelo de análise de sentimentos mais eficaz para classificar com precisão *tweets* inéditos. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

de NOGUEIRA (2021), foi incorporada ao processo. Dessa forma, a combinação dessas cinco APIs visou garantir maior precisão e robustez na classificação dos sentimentos.

Para submeter cada um dos *tweets* às APIs de análise de sentimentos, foi necessário criar uma conta de desenvolvedor em cada plataforma, para obtenção do acesso aos serviços disponibilizados. Após essa etapa, cinco *scripts* em python foram implementados, um para cada API. Esses *scripts* foram desenvolvidos para enviar cada *tweet* a cada uma das plataformas de análise de sentimentos e recuperar as classificações de sentimentos (negativa, neutra ou positiva) bem como as pontuações (*scores*) correspondentes.

4.1.4.6.2 Normalização dos dados

Cada ferramenta, conforme destacado por NOGUEIRA et al. (2024), apresenta um formato de dados específico. As APIs *Amazon Comprehend*, *GOTIT* e *Microsoft Text Analytics*, além de classificarem os sentimentos como negativos, neutros e positivos, também forneciam o rótulo *mixed*, indicando uma mistura de sentimentos. Para garantir uniformidade na análise, convencionou-se que todas as instâncias rotuladas como *mixed* seriam atribuídas à classe neutra, seguindo a mesma estratégia adotada por PINTO; ROCIO (2019).

Outro tratamento realizado foi a normalização das pontuações retornadas pelas APIs *Amazon Comprehend*, *Google Natural Language* e *Microsoft Azure*. O objetivo foi o de obter valores de pontuação entre -1 e 1. No entanto, cada ferramenta retorna uma em particular. Para normalizar os valores retornados pela *Amazon Comprehend* e *Microsoft Azure*, analisou-se em conjunto o rótulo e o *score*. Se o rótulo de uma determinada amostra fosse *NEGATIVE*, a pontuação seria multiplicada por -1; se fosse *NEUTRAL*, a pontuação normalizada seria 0; e se o rótulo fosse *POSITIVE*, a pontuação seria o próprio valor de pontuação sem qualquer manipulação. A ferramenta *Google Natural Language* retornava a pontuação como uma porcentagem. Para normalizar estes valores, multiplicou-se o valor retornado pela ferramenta por 0,01.

4.1.4.6.3 Abordagens propostas para a rotulagem automática do conjunto de dados de treinamento

Com os dados devidamente normalizados, 3 abordagens para rotulação automática dos dados foram implementadas. A primeira, denominada *OC Label Dominante*, utiliza uma estratégia de votação. Para cada *tweet*, analisou-se o rótulo retornado por cada uma das APIs, contabilizando o total de classificações para cada uma das classes negativa, neutra e positiva. A classe que obtivesse o maior valor era a classe que seria atribuída ao dado de nosso *dataset*. A segunda estratégia denominada *OC Score Médio* foi baseada no trabalho de PINTO; ROCIO (2019), em que se calculava a média aritmética dos *scores* de cada uma das APIs por meio da equação:

$$OC\ Score\ Médio = \frac{\sum_{i=1}^{n} score(i)}{n} \tag{4.1}$$

A terceira e última abordagem, denominada *OC Score Ponderado Médio*, também se baseia no trabalho de PINTO; ROCIO (2019). Como o próprio nome sugere, essa estratégia consiste no cálculo da média aritmética ponderada dos *scores*, em que os pesos correspondem ao número de classificações de mesma classe atribuídas pelas APIs a uma

determinada instância (*tweet*). A equação utilizada para essa metodologia é apresentada a seguir:

$$OC\ Score\ Ponderado\ Médio = \frac{\sum_{i=1}^{n} score(i) \cdot peso(i)}{\sum_{i=1}^{n} peso(i)} \tag{4.2}$$

4.1.4.6.4 Análise de concordância entre as APIs

Para verificar o quão as APIs concordavam entre si com relação a seus rótulos, utilizou-se o *Cohen Kappa Score*, o qual é uma função que calcula o nível de concordância entre dois anotadores em um problema de classificação. É definido como $\kappa = \frac{p_0 - p_\epsilon}{1 - p_\epsilon}$, em que p_0 é a probabilidade empírica de concordância no rótulo atribuído a qualquer amostra (a razão de concordância observada), e p_ϵ é a concordância esperada quando ambos os anotadores atribuem rótulos aleatoriamente (PEDREGOSA et al., 2011).

4.1.4.6.5 Abordagens para mitigação do desequilíbrio de dados

Uma característica presente em ambas versões do OC é o desbalanceamento entre as classes negativa, neutra e positiva. Para extrair os melhores resultados do modelo, aplicou-se a cada um dos *datasets* criados pelas abordagens *OC Label Dominante*, *OC Score Médio* e *OC Score Ponderado Médio* seis técnicas de reamostragem de dados. Utilizando o pacote *Imbalanced Learn* (LEMAîTRE et al., 2017) as técnicas *NearMiss*, *SMOTE*, *ADASYN*, *SMOTE-ENN* e *SMOTETomek* foram implementadas. Já a técnica *Back Translation* foi implementada utilizando a ferramenta *Marian*, responsável por realizar a tradução automática de textos de um idioma para outro (MARCIN et al., 2018). Os idiomas escolhidos para essa abordagem foram Português-Brasil (PT-BR) (idioma de origem) e o Inglês (EN) (idioma de destino), traduzindo ao fim deste processo novamente para Português-Brasil.

4.1.4.6.6 Implementação dos modelos de análise de sentimentos

Para esta implementação a linguagem de programação *Python* (PYTHON SOFTWARE FOUNDATION, 2021) foi utilizada em conjunto com as bibliotecas Scikit-Learn (PEDREGOSA et al., 2011), NLTK (LOPER; BIRD, 2002), Pandas (MCKINNEY, 2010), dentre outras. Para avaliar o efeito de cada uma das técnicas de balanceamento dos dados, o modelo de análise de sentimento implementado seguiu os mesmos princípios utilizados no trabalho de NOGUEIRA (2021) com algumas adaptações.

Em seu estudo, NOGUEIRA (2021) propõe um modelo que utiliza um esquema de votação denominado *Ensemble Voting Classifier*. As abordagens de *ensemble* são responsáveis por combinar vários classificadores de aprendizagem de máquina para predição dos rótulos das classes. Ao utilizar este esquema de votação, seja por meio da maioria de

votos ou por meio de probabilidades médias (PEDREGOSA et al., 2011) falhas individuais dos classificadores são compensadas e minimizadas devido à consideração da opinião dos demais classificadores. Este fator impacta diretamente no desempenho final do modelo com a atenuação da perda de desempenho.

Neste trabalho os classificadores utilizados no *Voting Classifier* foram: Regressão Logística, *Multinomial Naive Bayes*, *OneVsRestClassifier*, *OneVsOneClassifier*, *KNN*, *Random Forest*, *SVM*, e *Linear SVC*. NOGUEIRA (2021) empregou os quatro primeiros em sua pesquisa, enquanto os demais foram utilizados conforme as implementações descritas por HNAIF et al. (2021) e SAURA et al. (2021).

Para validar o modelo, foi empregada a técnica de validação cruzada, cujo objetivo é avaliar a capacidade de generalização do modelo ao medir sua precisão diante de novos conjuntos de dados. Utilizando o *K-fold* o conjunto de dados de treinamento foi dividido em k partes, uma para treinamento e a outra para teste.

Especificamente, foi adotada a função *StratifiedKFold* com 10 *folds* e o parâmetro *shuffle = True*, garantindo o embaralhamento dos dados e a manutenção da proporção de cada classe em todas as divisões. Essa abordagem permite uma distribuição equilibrada das classes entre as partições, reduzindo os impactos do desbalanceamento da base de treinamento. Dessa forma, assegura-se que cada *fold* seja representativo do conjunto original, contribuindo para uma avaliação mais robusta e confiável do desempenho do modelo.

Para cada um dos conjuntos de dados de treinamento construídos pelas abordagens *OC Label Dominante*, *OC Score Médio* e *OC Score Ponderado Médio*, os algoritmos de balanceamento de classes foram aplicados para reajustá-los e submetê-los ao final desta etapa ao modelo de análise de sentimentos. Para comparar os resultados de cada modelo construído, quatro principais métricas foram analisadas (i) *precision*, (ii) *recall*, (iii) *F1-score* e (iv) *support* contidas no relatório da matriz de confusão.

- *precision*: é a razão $\frac{tp}{tp+fp}$ em que tp é o número de verdadeiros positivos e fp o número de falsos positivos. A precisão é intuitivamente a capacidade do classificador não rotular como positiva uma amostra que é negativa;

- *recall*: é a razão $\frac{tp}{tp+fn}$ em que tp é o número de verdadeiros positivos e fn o número de falsos negativos. O *recall* pode ser definido como a capacidade do classificador encontrar todas as amostras positivas;

- *F1-score*: A pontuação de F1 pode ser interpretada como uma média harmônica da precisão e *recall*, em que uma pontuação de F1 atinge seu melhor valor em 1 e pior pontuação em 0. A contribuição relativa da precisão e *recall* para a pontuação de F1 é igual. A fórmula para esta pontuação é dada por: $F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$;

- *Support*: é o número de ocorrências de cada classe em y_true.

4.1.4.6.7 Análise da capacidade de generalização dos modelos de análise de sentimentos

Para avaliar a capacidade de generalização dos modelos desenvolvidos, foi utilizado um conjunto de teste composto por 10.585 *tweets*, rotulados seguindo o mesmo processo aplicado ao conjunto de treinamento. Segundo NOGUEIRA et al. (2024), esse conjunto era distribuído em 2.892 *tweets* negativos, 3.990 neutros e 3.703 positivos.

Os mesmos autores destacaram que essas amostras de teste foram mantidas completamente separadas do treinamento, garantindo uma avaliação justa e imparcial do desempenho dos modelos em dados inéditos. A análise dos resultados nesse conjunto permitiu não apenas a comparação entre as diferentes versões dos modelos desenvolvidos, mas também uma melhor compreensão de sua capacidade de classificar corretamente sentimentos em novos dados, assegurando a robustez das soluções propostas.

4.1.4.7 Análise de consumidores e potenciais consumidores

A análise de consumidores e potenciais consumidores, inicialmente implementada por NOGUEIRA (2021), foi revisitada e aprimorada na nova versão do OC. Na versão anterior, a identificação de padrões de consumo era baseada em um conjunto restrito de verbos, sem considerar a relação sintática das publicações. Essa limitação resultava em imprecisões, pois um verbo poderia estar associado a produtos consumidos em conjunto com os lácteos, sem indicar necessariamente o consumo direto dos produtos lácteos de interesse.

Na nova abordagem, a identificação do consumo efetivo foi refinada por meio da análise de verbos pertencentes à categoria Consumo de Alimentos, levando em conta sua flexão nos tempos verbais passado e presente, com o auxílio da biblioteca *spaCy*. Além disso, para identificar possíveis intenções de consumo futuro, foram analisadas flexões verbais no tempo futuro, bem como a presença de verbos auxiliares que pudessem indicar essa intenção nas publicações. Essas melhorias garantem uma abordagem mais precisa e contextualizada na detecção de padrões de consumo.

4.1.5 Pós-processamento dos dados e visualização dos resultados

Após a execução das etapas anteriores, é possível analisar e correlacionar os resultados obtidos, o que permite identificar informações relevantes e obter uma compreensão mais aprofundada do setor. Com os dados organizados, inicia-se o processo de relacioná-los para extrair informações, capazes de responder a questões que as ferramentas tradicionais de BI abordam, como: “Qual produto foi consumido?”, “Onde ocorreu o consumo?” (localização), “Quando ocorreu o consumo?” (dia e período), “Qual foi o sentimento associado

ao consumo?" (negativo, neutro ou positivo), "Quem consumiu?" (homem ou mulher), "A qual geração pertence o consumidor?" (geração silenciosa, baby boomers, X, Y, Z ou alfa), "Por que consumiu?" (motivos expressos por verbos), "Como foi o consumo?" (descrições em adjetivos) e "Com o que foi consumido?" (hábitos de consumo).

Essas questões são abordadas por meio das análises oferecidas pelo OC, com cada rede social contribuindo conforme os dados que disponibiliza. Nesse contexto, a Tabela 4 apresenta um resumo comparativo das análises realizadas pelo OC para cada fonte de dados empregada.

Tabela 4 – Comparação das análises realizadas por cada fonte de dados utilizada pelo OC. **Fonte:** autor.

Pergunta que a análise responde	**Análises realizadas**	**Redes Sociais**		
		Plataforma X	**YouTube**	**Google Trends**
Quem?	Análise do Sexo	X	-	-
	Análise da Geração	X	-	-
	Análise de Renda	X	-	-
Quando?	Análise Temporal	X	X	X
Qual Produto?	Análise de Produtos Lácteos Mencionados	X	X	X
Com o Quê?	Análise dos demais Produtos Mencionados	X	X	X
Onde?	Análise Geográfica	X	-	X
Como foi?	Análise Comportamental	X	X	-
	Análise de Sentimentos	X	X	-
	Características do Produto	X	X	-
	Experiência do Consumidor	X	X	-

As análises realizadas pelo OC respondem a essas questões, permitindo identificar tendências, características e padrões de consumo dos produtos lácteos no Brasil, contribuindo para uma análise mais estratégica e direcionada do mercado. Para tornar o grande volume de informações extraídas mais acessível, o OC foi projetado como uma ferramenta simples e intuitiva. Assim, o sistema permite que os usuários realizem consultas próprias e obtenham conclusões a partir dos dados disponibilizados pela plataforma.

A primeira versão do OC, por se tratar de um MPV com baixo volume de dados, utilizou a linguagem PHP sem *frameworks* adicionais. Diante desse cenário, o primeiro passo para a reformulação do OC foi a escolha de um *framework* robusto o suficiente para atender às novas demandas do projeto. Optou-se, então, pelo *Laravel* (OTWELL, 2011), uma solução de código aberto baseada na linguagem PHP (BAKKEN et al., 2000) que adota o padrão de arquitetura MVC (*Model-View-Controller*), que organiza e separa as responsabilidades do código em três componentes principais: o modelo, a visão e o controlador. Esse padrão ajuda a manter o código limpo, modular e fácil de manter (VICENTE; SIRQUEIRA, 2022).

Como já mencionado, o banco de dados utilizado na ferramenta foi o MySQL (MYSQL AB, 2024), enquanto o controle de versões do sistema foi realizado via GitHub (GITHUB, 2024), permitindo que os desenvolvedores trabalhassem de forma independente. O sistema foi hospedado em um servidor *Ubuntu*, disponibilizado pela Embrapa Gado de Leite. Com essa arquitetura computacional, diversos trabalhos foram implementados, os quais serão brevemente descritos no próximo capítulo 5, intitulado “Análise dos Resultados e Discussão”.

5 ANÁLISE DOS RESULTADOS E DISCUSSÃO

Este capítulo apresenta em detalhes os principais resultados obtidos com a aplicação do Observatório do Consumidor, enfatizando sua influência na produção científica. Serão analisadas algumas publicações científicas, bem como os impactos da ferramenta na condução dessas pesquisas e os avanços proporcionados por seu uso.

5.1 Trabalho "*Construction of a Training Dataset for a Sentiment Analysis Model of Dairy Products Tweets in Brazil*"

Esta seção apresenta os resultados do trabalho de NOGUEIRA et al. (2024), que teve como principal objetivo desenvolver e implementar um conjunto de dados de treinamento equilibrado e representativo para um modelo de análise de sentimentos focado no contexto do mercado de produtos lácteos no Brasil.

5.1.1 Resultados da etapa de rotulação automática dos dados

Conforme descrito na subseção 4.1.4.6.3, o estudo de NOGUEIRA et al. (2024) propôs três abordagens de rotulação automática: *OC Label Dominante*, *OC Score Médio* e *OC Score Ponderado Médio*. Essas estratégias foram aplicadas na construção dos conjuntos de dados para treinamento, validação e teste dos modelos de análise de sentimentos, sendo desenvolvidas a partir da integração dos resultados gerados por cada uma das APIs utilizadas. Após a submissão das 86.060 amostras dos conjuntos de treinamento e validação às APIs, e a recuperação dos resultados fornecidos por elas, a Figura 19 apresenta uma compilação desses dados.

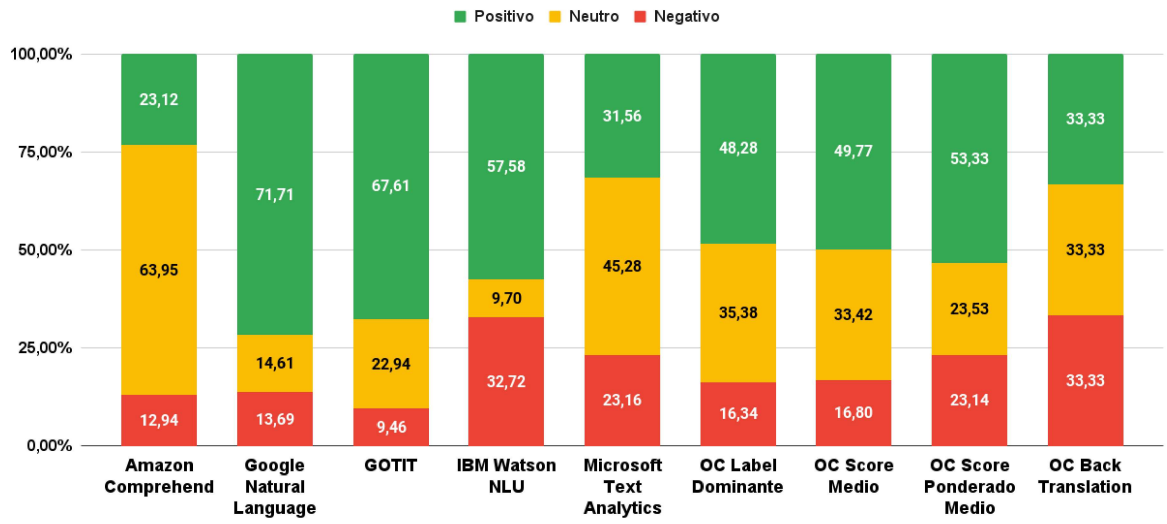


Figura 19 – Representatividade de classes por API e abordagens utilizadas. **Fonte:** autor.

É possível observar uma significativa variabilidade nas classificações obtidas pelas APIs analisadas. Um dos fatores que podem ter influenciado esse resultado é a origem dos dados. Como destacado por SILVA (2016), tarefas que envolvem PLN são intrinsecamente complexas, especialmente quando aplicadas a dados provenientes de redes sociais. Essa complexidade é amplificada por desafios linguísticos, como a ambiguidade de palavras e frases, o uso de sarcasmo, ironia, gírias, erros ortográficos, regionalismos e dialetos, que dificultam a categorização precisa da polaridade dos sentimentos. No entanto, essa diversidade de resultados ressalta a importância da abordagem adotada por NOGUEIRA et al. (2024), que buscou consolidar todas as informações disponíveis para a criação de modelo que agregasse efetivamente os resultados de cada ferramenta, garantindo, assim, uma resposta mais precisa, consistente e confiável.

Mesmo com um número ímpar de APIs utilizadas (cinco no total), ocorreram casos de empate na abordagem *OC Label Dominante*. O conjunto de dados apresentou 12.798 casos de empate, representando 14,87% do total. Desses, 3.450 empates ocorreram entre as classes negativa e neutra, 3.170 entre negativa e positiva, e 6.178 entre neutra e positiva.

Para resolver os casos de empate, adotou-se a abordagem do *OC Score Médio*. Essa escolha foi motivada pelo fato de que, ao contrário do *OC Score Ponderado Médio*, essa metodologia não considerava os pesos atribuídos às pontuações, baseando-se exclusivamente nos valores retornados por cada API. Para a análise, foram selecionados todos os *tweets* envolvidos nos empates, juntamente com suas respectivas pontuações. Em seguida, essas pontuações foram organizadas para gerar curvas de distribuição associadas a cada classe. Esse processo foi essencial para determinar os intervalos de valores que melhor representam cada uma das classes. A Figura 20 ilustra a distribuição dos *scores* obtidos por meio da abordagem do *OC Score Médio*, fornecendo uma visualização clara dos padrões identificados.

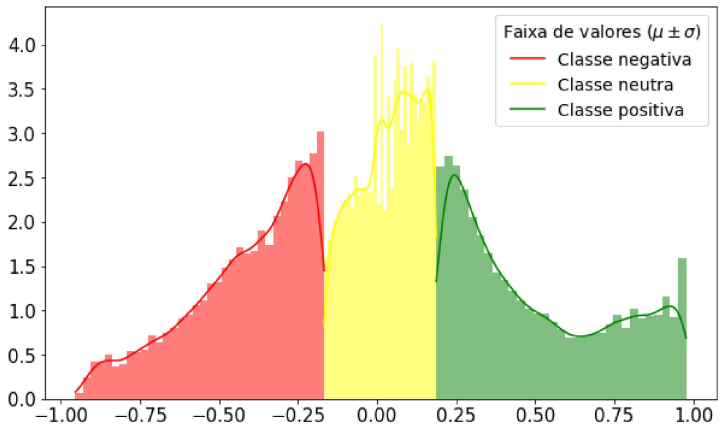


Figura 20 – Distribuição dos *scores* para a abordagem *OC Score Médio* por classe. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Durante a análise das distribuições dos *scores*, NOGUEIRA et al. (2024) estabele-

ceram que a classe neutra seria atribuída a *tweets* cujos valores de *score* estivessem no intervalo $[-0,16645; 0,18780]$. Essa definição foi baseada na observação de que, à medida que os valores se afastavam desse intervalo, aumentava a probabilidade de classificação incorreta de *tweets* pertencentes às classes negativa ou positiva, o que poderia comprometer a precisão dos resultados. Após a aplicação desse critério para resolver os casos de empate, os resultados consolidados para cada conjunto de dados foram organizados e estão detalhados na Tabela 5.

Tabela 5 – Número de amostras por classe no conjunto de treinamento antes do balanceamento. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

OC Label Dominante e OC Back Translation			**OC Score Médio**			**OC Score Ponderado Médio**		
Neg	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**
14.065	30.444	41.551	14.462	28.765	42.833	19.914	20.252	45.894

A Tabela 5 mostra que o conjunto de dados *OC Label Dominante* serviu como base para a criação do conjunto de dados *OC Back Translation*. A técnica de *Back Translation* foi aplicada de forma estratégica, focando exclusivamente na classe negativa, uma vez que essa era a classe minoritária, com o menor número de amostras no conjunto original. NOGUEIRA et al. (2024) mencionam que a decisão de concentrar esforços nessa classe específica foi tomada devido ao elevado custo computacional associado ao processo de tradução de textos entre idiomas, o que tornaria inviável sua aplicação em larga escala para todas as classes.

Além disso, o desequilíbrio significativo observado em todas as abordagens de rotulação ressalta a importância de técnicas de balanceamento de dados. Essas técnicas são fundamentais para assegurar a representatividade adequada das classes, especialmente em cenários em que uma classe é substancialmente mais frequente que as outras. Ao equilibrar os dados, é possível mitigar vieses que poderiam comprometer a eficácia dos modelos de análise de sentimentos, resultando em previsões mais precisas e confiáveis.

5.1.2 Análise comparativa da classificação de sentimentos: APIs versus abordagens propostas

Conforme ilustrado na Figura 19, cada API utilizada para a pré-rotulação dos dados atribuiu classificações de polaridade distintas aos *tweets* nos conjuntos analisados. Essa variabilidade nas classificações destacou a necessidade de uma análise mais aprofundada sobre a consistência entre as APIs. Para isso, NOGUEIRA et al. (2024) fizeram uma análise mais detalhada por meio da matriz de coeficientes *Kappa*, apresentada na Figura 21. Essa matriz possibilita a avaliação do nível de concordância entre as APIs, oferecendo uma visão clara da consistência e da confiabilidade das classificações realizadas por cada ferramenta.

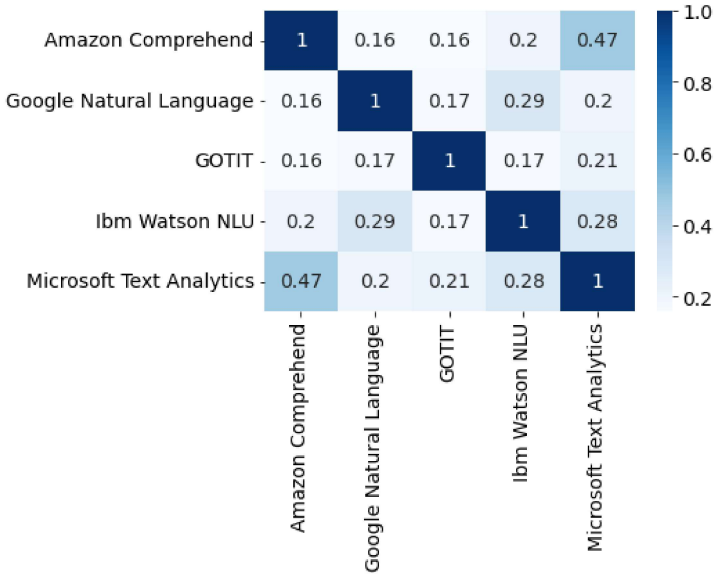


Figura 21 – Concordância dos resultados das APIs utilizando os coeficientes de *kappa*. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

É possível constatar uma discordância significativa entre as classificações fornecidas pelas diferentes APIs. A análise da Figura 21 revela que, ao comparar a ferramenta *Amazon Comprehend* com as demais, a *Microsoft Text Analytics* apresentou o maior nível de concordância, alcançando 46,7%. Embora esse valor seja modesto, representa o melhor resultado observado. Em seguida, tanto o *IBM Watson NLU* quanto o *Google Natural Language* obtiveram um índice de concordância de 28,7%, ocupando a segunda posição no *ranking* indicando o segundo maior nível de concordância entre as APIs. Esses resultados reforçaram a relevância da análise de NOGUEIRA et al. (2024), uma vez que a utilização isolada de qualquer uma dessas ferramentas não assegurava uma classificação confiável e precisa dos dados, o que poderia comprometer a generalização adequada das classes de sentimento pelo modelo.

Diante disto, a metodologia proposta por NOGUEIRA et al. (2024) visava integrar os resultados obtidos de cada API, reforçando a confiabilidade das classificações ao consolidar as análises de cinco ferramentas distintas, em vez de depender de uma única fonte. Nesse contexto, a Figura 22 apresenta uma comparação detalhada do nível de concordância entre as abordagens *OC Label Dominante*, *OC Score Médio* e *OC Score Ponderado Médio* e as demais ferramentas utilizadas.

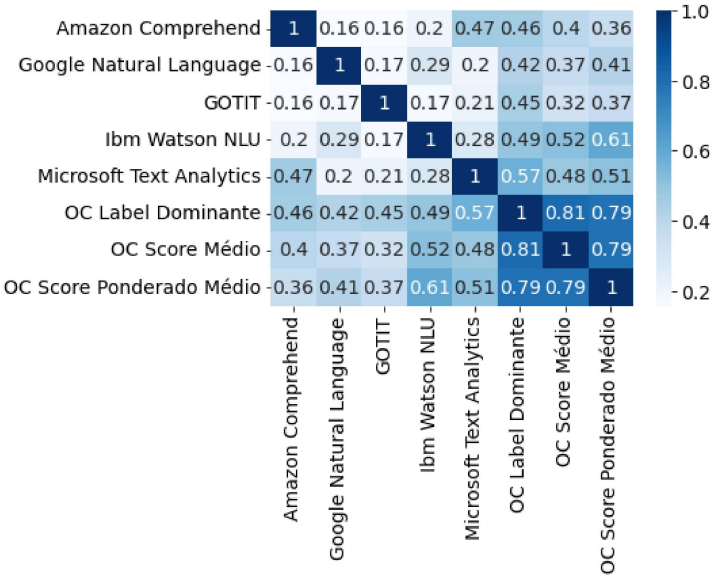


Figura 22 – Comparação da concordância dos resultados das APIs com as abordagens propostas. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Conforme ilustrado na Figura 22, a concordância entre os métodos propostos é notavelmente mais robusta, com valores de 0,79 e 0,81. A comparação dos métodos propostos com os resultados individuais das APIs revela uma melhoria significativa, embora a concordância ainda seja moderada. Esse resultado destaca a importância de integrar as pontuações de cada API, o que desempenhou um papel fundamental no desenvolvimento dos conjuntos de dados de treinamento. Ao incorporar as características de todas as cinco ferramentas de classificação, criamos conjuntos de dados mais abrangentes e representativos, resultando em uma classificação com maior confiança.

5.1.3 Resultados obtidos com a aplicação de métodos de balanceamento de dados nos conjuntos de treinamento

Os métodos de balanceamento de dados são cruciais na preparação de conjuntos de dados para análises de aprendizado de máquina. Cada técnica possui características distintas que influenciam na distribuição das amostras entre as classes. A Tabela 6 apresenta os resultados da aplicação desses métodos nos conjuntos de dados *OC Label Dominante*, *OC Score Médio* e *OC Score Ponderado Médio*.

Tabela 6 – Número de amostras por classe no conjunto de dados de treinamento após o balanceamento. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Métodos de balanceamento	*OC Label Dominante*			*OC Score Médio*			*OC Score Médio Ponderado*		
	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	Pos
NearMiss	14.065	14.065	14.065	14.462	14.462	14.462	19.914	19.914	19.914
SMOTE	41.551	41.551	41.551	42.833	42.833	42.833	45.894	45.894	45.894
ADASYN	41.343	40.849	41.551	43.664	53.326	42.833	41.980	39.782	45.894
SMOTE-ENN	30.822	13.403	2.391	30.892	11.608	2.953	22.742	23.787	2.880
SMOTETomek	41.527	41.086	41.098	42.786	42.424	42.437	45.800	45.679	45.675

Os resultados encontrados por NOGUEIRA et al. (2024) indicaram que as técnicas *NearMiss* e *SMOTE* alcançaram uma distribuição equilibrada entre as classes, assegurando um número igual de amostras para cada classe em todos os conjuntos de dados. Em contraste, o método *ADASYN* apresentou uma distribuição desigual, sendo que o conjunto de dados *OC Label Dominante* exibiu a distribuição mais equilibrada em relação aos demais. Já o método *SMOTE-ENN* gerou um número significativamente maior de amostras para a classe negativa em comparação às classes neutra e positiva, resultando em uma distribuição desequilibrada. Por outro lado, o *SMOTETomek* conseguiu equilibrar as classes, embora tenha apresentado pequenas variações no número de amostras entre elas.

No caso do conjunto de dados *OC Back Translation*, inicialmente não foi adequado incluir seus resultados na Tabela 6, pois essa técnica utiliza exclusivamente os dados do conjunto *OC Label Dominante*. A técnica de *Back Translation* equilibrava os dados da classe negativa por meio de duas etapas distintas. Na primeira etapa, a técnica foi aplicada aos dados originais da classe negativa, gerando novas instâncias. Em seguida, o procedimento foi repetido nos dados recém-gerados, introduzindo ainda mais variabilidade. Esse processo iterativo resultou na criação de 11.632 novos dados na primeira rodada e 7.057 na segunda rodada. Após a remoção de duplicatas, o conjunto final da classe negativa passou a conter 30.442 instâncias.

Ao aplicar a técnica de *Back Translation*, era esperado que as publicações mantivessem sua polaridade original após serem traduzidas para outro idioma e, posteriormente, traduzidas de volta para o idioma original. Contudo, essa suposição revelou-se inválida em alguns casos. Durante o processo de validação da integridade dos dados, NOGUEIRA et al. (2024) identificaram 6.610 *tweets* recém-gerados cujas polaridades apresentaram divergências após a retro tradução. Essas instâncias foram removidas para assegurar a confiabilidade do conjunto de dados. Como resultado desse procedimento, um conjunto de dados equilibrado foi obtido, com 23.832 *tweets* classificados como negativos, e um número equivalente de amostras para as classes neutra e positiva. Assim, o conjunto de treinamento *OC Back Translation* foi composto por 71.496 instâncias, distribuídas uniformemente entre as três classes.

Os resultados obtidos por NOGUEIRA et al. (2024) forneceram informações importantes sobre os efeitos dos diferentes métodos de balanceamento nos conjuntos de dados analisados. Esses métodos desempenharam um papel fundamental ao melhorar a representatividade e o equilíbrio das amostras, otimizando os conjuntos de treinamento para análises mais precisas, confiáveis e com maior aplicabilidade prática.

5.1.4 Análise do desempenho dos modelos de análise de sentimentos com técnicas de balanceamento de dados

Após a preparação dos conjuntos de dados, NOGUEIRA et al. (2024) realizaram a avaliação desses conjuntos nos modelos de análise de sentimentos implementados. O objetivo principal foi identificar a abordagem de rotulação automática mais eficaz, o modelo de análise de sentimentos mais representativo e a técnica de balanceamento mais adequada para os conjuntos de dados construídos.

NOGUEIRA et al. (2024) justificaram a não adoção de modelos baseados em *deep learning* por motivos específicos. Em primeiro lugar, os experimentos demonstraram que as técnicas de balanceamento de dados aumentaram consideravelmente os custos computacionais, exigindo maior capacidade de processamento. Embora os modelos de *deep learning* apresentassem potencial para resultados mais robustos, os autores destacaram que os recursos computacionais necessários tornaram sua aplicação inviável. Por isso, optaram por métodos tradicionais, que além de proporcionarem resultados satisfatórios, demandam significativamente menos poder computacional. Essa escolha promoveu maior acessibilidade, eficiência e sustentabilidade, alinhando-se às limitações práticas e aos objetivos da pesquisa. Os resultados obtidos estão apresentados na Tabela 7.

Tabela 7 – Resultados dos modelos de análise de sentimentos aplicados às diferentes técnicas de balanceamento de dados e aos conjuntos de dados de treinamento propostos. Legenda: AM - Acurácia Média (%), IC - Intervalo de Confiança (%). **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Técnica de Balanceamento	*OC Label Dominante*		*OC Score Médio*		*OC Score Ponderado Médio*	
	AM	IC	AM	IC	AM	IC
Sem balanceamento	71,45	[70,80 - 72,10]	71,24	[70,38 - 72,10]	71,77	[71,06 - 72,47]
NearMiss	70,51	[68,68 - 72,34]	68,87	[67,27 - 70,46]	68,04	[66,41 - 69,66]
SMOTE	79,98	[79,64 - 80,32]	78,40	[77,77 - 79,03]	81,95	[81,26 - 82,65]
ADASYN	78,54	[77,84 - 79,25]	82,68	[82,25 - 83,10]	80,59	[79,79 - 81,40]
SMOTE-ENN	**92,00**	**[91,61 - 92,39]**	**92,02**	**[91,31 - 92,72]**	**93,90**	**[93,32 - 94,47]**
SMOTETomek	80,14	[79,60 - 80,68]	78,62	[77,96 - 79,28]	82,28	[81,83 - 82,73]
Back Translation	72,22	[70,99 - 73,44]	-	-	-	-

O melhor resultado para a acurácia média está destacado em negrito.

Os efeitos positivos das técnicas de balanceamento de dados são evidentes, embora a técnica *NearMiss* tenha apresentado resultados inferiores em comparação com as demais abordagens propostas. O melhor desempenho dessa técnica foi observado ao aplicá-la ao conjunto de dados *OC Label Dominante* durante o treinamento. Esse resultado pode ser explicado pela composição do *OC Label Dominante*, que apresenta uma menor proporção de amostras da classe negativa em relação aos outros dois conjuntos. Como consequência, o tamanho final do conjunto de dados balanceado é reduzido, o que pode ter impactado positivamente a acurácia da classificação. Para uma análise detalhada das métricas de *precision*, *recall*, *F1-Score* e *support*, consulte a Tabela 8.

Tabela 8 – Resultados das métricas de avaliação nas etapas de treinamento e validação do modelo utilizando diferentes técnicas de balanceamento de dados nos conjuntos de dados propostos por cada abordagem de rotulagem. Legenda: OC I - *Label Dominante*, OC II - *Score Médio*, OC III - *Score Ponderado Médio*. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Modelos criados com as respectivas técnicas de balanceamento de dados	Métricas	Dataset	Polaridade dos Sentimentos		
			Negativo	Neutro	Positivo
Sem balanceamento	*Precision*	OC I	0,78	0,63	0,77
		OC II	0,77	0,60	0,77
		OC III	0,74	0,57	0,75
	Recall	OC I	0,50	0,70	0,79
		OC II	0,51	0,65	0,82
		OC III	0,63	0,45	0,88
	F1-Score	OC I	0,61	0,66	0,78
		OC II	0,61	0,62	0,80
		OC III	0,68	0,50	0,81
	Support	OC I	14.065	30.444	41.551
		OC II	14.462	28.765	42.833
		OC III	19.914	20.252	45.894
NearMiss	*Precision*	OC I	0,71	0,65	0,76
		OC II	0,70	0,62	0,76
		OC III	0,68	0,61	0,76
	Recall	OC I	0,81	0,64	0,67
		OC II	0,80	0,58	0,69
		OC III	0,77	0,62	0,66
	F1-Score	OC I	0,76	0,64	0,71
		OC II	0,74	0,60	0,72
		OC III	0,72	0,62	0,71

Continua na próxima página

Tabela 8 – *Continuação da tabela*

Modelos criados com as respectivas técnicas de balanceamento de dados	Métricas	Dataset	Polaridade dos Sentimentos		
			Negativo	Neutro	Positivo
	Support	OC I	14.065	14.065	14.065
		OC II	14.462	14.462	14.462
		OC III	19.914	19.914	19.914
SMOTE	*Precision*	OC I	0,81	0,75	0,85
		OC II	0,77	0,73	0,85
		OC III	0,82	0,77	0,87
	Recall	OC I	0,95	0,73	0,71
		OC II	0,95	0,67	0,74
		OC III	0,91	0,81	0,73
	F1-Score	OC I	0,87	0,74	0,78
		OC II	0,85	0,70	0,79
		OC III	0,86	0,79	0,80
	Support	OC I	41.551	41.551	41.551
		OC II	42.833	42.833	42.833
		OC III	45.894	45.894	45.894
ADASYN	*Precision*	OC I	0,79	0,73	0,85
		OC II	0,86	0,76	0,90
		OC III	0,80	0,76	0,86
	Recall	OC I	0,96	0,70	0,70
		OC II	0,94	0,85	0,80
		OC III	0,90	0,75	0,76
	F1-Score	OC I	0,86	0,71	0,76
		OC II	0,90	0,80	0,77
		OC III	0,85	0,76	0,81
	Support	OC I	41.434	40.849	41.551
		OC II	43.664	53.326	42.833
		OC III	41.980	39.782	45.894
SMOTE-ENN	*Precision*	OC I	0,90	0,99	0,99
		OC II	0,90	0,97	0,99
		OC III	0,93	0,95	0,95
	Recall	OC I	0,99	0,76	0,84
		OC II	1,00	0,72	0,89
		OC III	0,96	0,93	0,84
	F1-Score	OC I	0,94	0,85	0,91
		OC II	0,95	0,83	0,94

Continua na próxima página

Tabela 8 – *Continuação da tabela*

Modelos criados com as respectivas técnicas de balanceamento de dados	Métricas	Dataset	Polaridade dos Sentimentos		
			Negativo	Neutro	Positivo
		OC III	0,95	0,94	0,89
	Support	OC I	30.822	13.403	2.391
		OC II	30.892	11.608	2.953
		OC III	22.742	23.787	2.880
SMOTETomek	*Precision*	OC I	0,81	0,75	0,85
		OC II	0,77	0,74	0,85
		OC III	0,82	0,78	0,88
	Recall	OC I	0,95	0,73	0,72
		OC II	0,95	0,67	0,74
		OC III	0,91	0,82	0,74
	F1-Score	OC I	0,88	0,74	0,78
		OC II	0,85	0,70	0,79
		OC III	0,86	0,80	0,80
	Support	OC I	41.527	41.086	41.098
		OC II	42.786	42.424	42.437
		OC III	45.800	45.679	45.675
Back Translation	*Precision*	OC I	0,75	0,65	0,78
		OC II	-	-	-
		OC III	-	-	-
	Recall	OC I	0,83	0,65	0,69
		OC II	-	-	-
		OC III	-	-	-
	F1-Score	OC I	0,79	0,65	0,73
		OC II	-	-	-
		OC III	-	-	-
	Support	OC I	23.832	23.832	23.832
		OC II	-	-	-
		OC III	-	-	-

Fim da tabela

Ao analisar os resultados, NOGUEIRA et al. (2024) observaram que a discrepância no número de amostras entre os modelos evidenciou a necessidade de investigações adicionais para compreender melhor os impactos dessas diferenças na capacidade de generalização dos modelos. Em outras palavras, é essencial entender como essas variações influenciam a habilidade dos modelos em classificar dados desconhecidos, ou seja, dados que não foram apresentados durante as fases de treinamento e validação. Para avaliar essa capacidade

de generalização, os modelos foram testados utilizando o conjunto de dados de teste. Os resultados dessa avaliação estão detalhados na Tabela 9.

Tabela 9 – Resultados dos modelos de análise de sentimentos no conjunto de teste. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Técnica de balanceamento dos dados	Acurácia Média (%)		
	OC Label Dominante	*OC Score Médio*	*OC Score Ponderado Médio*
Sem balanceamento	75.75	71.23	71.76
NearMiss	70.99	68.65	74.76
SMOTE	77.30	73.09	78.58
ADASYN	74.23	75.45	77.97
SMOTE-ENN	45.77*	42.05*	55.87*
SMOTETomek	77.17	72.94	**79.61**
Back Translation	77.24	-	-

O melhor resultado para a acurácia média está destacado em negrito, enquanto os dados marcados com um asterisco indicam os piores resultados identificados.

Os resultados apresentam diversos pontos de destaque, sendo um dos mais significativos o impacto da técnica *SMOTE-ENN* na capacidade de generalização dos modelos em relação aos dados da classe positiva. Conforme observado na Tabela 7, durante a fase de treinamento, o modelo que utilizou essa técnica alcançou os melhores resultados em termos de acurácia média. Contudo, ao ser testado com dados desconhecidos, o desempenho do modelo mostrou-se insatisfatório, ficando aquém até mesmo daquele que não aplicou nenhuma técnica de balanceamento de dados.

Por outro lado, o modelo *SMOTETomek* merece destaque por apresentar o melhor desempenho em termos de acurácia média na fase de teste, alcançando 79,61%. Esse valor está em estreita proximidade com a acurácia média de 82,28% obtida durante a fase de treinamento, ambos aplicados ao conjunto de dados *OC Score Ponderado Médio*. Essa consistência entre as fases de treinamento e teste reforçam a eficácia da técnica *SMOTETomek* em aprimorar o desempenho do modelo. Além disso, evidencia uma capacidade mais robusta de generalização, permitindo que o modelo classifique corretamente as classes de polaridade de sentimentos de postagens inéditas, nunca antes apresentadas durante o treinamento.

Ao comparar os resultados deste estudo com a literatura, foi constatado que os métodos de balanceamento de dados apresentaram desempenhos consistentes com os relatados em pesquisas voltadas para diferentes aplicações. No trabalho de CAMACHO (2020), que investigou o desenvolvimento de um modelo híbrido de recomendação e classificação, tanto o *ADASYN* quanto o *SMOTETomek* demonstraram resultados destacados, alinhando-se às descobertas deste estudo. Por sua vez, no estudo de SRIDHAR; SANAGAVARAPU (2021), cujo foco foi a distinção entre falhas e não falhas em máquinas, o método *SMOTE*

proporcionou um aumento de 7,83% no valor do AUC, melhorando significativamente o desempenho do classificador. Esses resultados reafirmam a eficácia e a adaptabilidade dos métodos de balanceamento de dados em diferentes contextos e aplicações.

Para uma análise mais detalhada dos resultados obtidos na fase de teste, os desempenhos foram avaliados individualmente para cada classe: negativa, neutra e positiva. É fundamental revisitar as características do conjunto de dados de teste previamente descritas por NOGUEIRA et al. (2024) para contextualizar esses resultados. O conjunto de teste era composto por 2.892 *tweets* na classe negativa, 3.990 na classe neutra e 3.703 na classe positiva. A Figura 23 apresenta uma comparação detalhada entre os diferentes conjuntos de dados criados, destacando o desempenho do modelo ao lidar com os dados de teste. Essa comparação foi ilustrada por meio de um mapa de calor, que demonstra a porcentagem da diferença absoluta entre o número real de amostras em cada classe e as previsões feitas pelo modelo, permitindo uma visualização clara do desempenho em termos de precisão por classe.

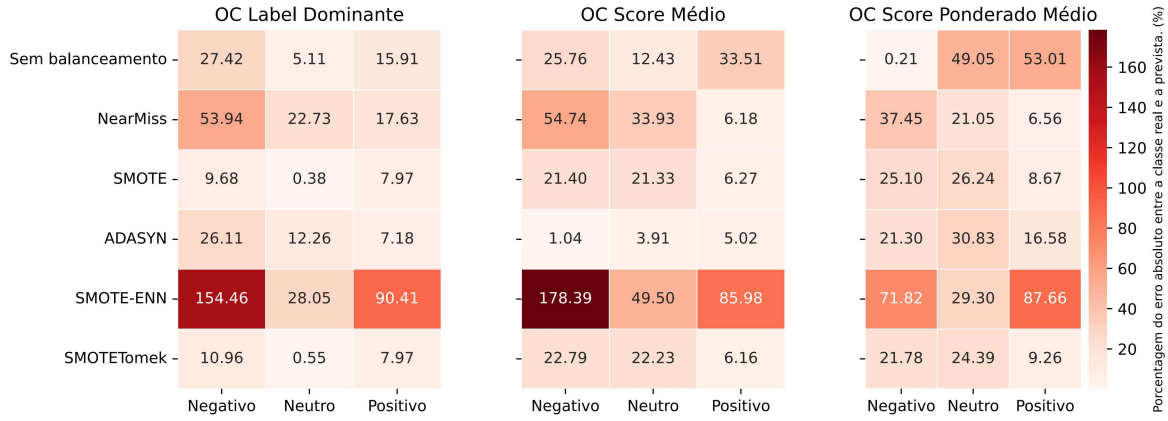


Figura 23 – Resultado do classificador do modelo de análise de sentimentos no conjunto de dados de teste, apresentado em um mapa de calor por classe, ilustra a porcentagem da diferença absoluta entre o total real de amostras em cada classe e as amostras previstas pelo modelo. Valores menores e cores mais claras indicam melhor desempenho, ou seja, maior precisão. Para evitar problemas de visualização no mapa de calor, não incluímos os resultados da técnica de *Back Translation*, pois ela foi aplicada apenas ao conjunto de dados *OC Label Dominante*. No entanto, seus resultados foram 6,94% para a classe negativa, 4,84% para a classe neutra e 10,61% para a classe positiva. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Novamente, os modelos que utilizaram a técnica de balanceamento de dados *SMOTE-ENN* apresentaram o pior desempenho. Esse resultado indica que a aplicação dessa técnica na tarefa de classificação de polaridade de sentimentos em postagens não se mostrou tão eficaz quanto observado durante a fase de treinamento. Considerando o objetivo principal de classificar dados não vistos pelo modelo, é evidente que os modelos que utilizaram esse método não atingiram resultados satisfatórios, sugerindo um possível caso de *overfitting*. O *overfitting* ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, comprometendo sua capacidade de generalizar para novos dados não observados, o que pode explicar o desempenho inferior observado na fase de teste.

Os resultados variaram conforme o conjunto de dados analisado. No caso do conjunto *OC Label Dominante*, NOGUEIRA et al. (2024) observaram que a aplicação das técnicas *SMOTE* e *SMOTETomek* resultou nas menores porcentagens de erro, indicando que esses métodos de balanceamento proporcionaram os melhores desempenhos na classificação de dados não vistos anteriormente. Para o conjunto *OC Score Médio*, o método *ADASYN* se destacou como o mais eficiente, superando os outros métodos de balanceamento de dados e conjuntos de dados. Importante ressaltar que essa técnica obteve a menor porcentagem de erro na classe negativa, sugerindo que o balanceamento realizado foi especialmente eficaz para essa classe minoritária no conjunto original. Quanto ao conjunto de dados *OC Score Ponderado Médio*, os resultados foram menos expressivos do que nos dois conjuntos anteriores, embora a porcentagem de erro absoluto tenha se mostrado semelhante à observada nos demais conjuntos.

5.1.5 Análise qualitativa dos *tweets* classificados incorretamente pelo melhor modelo no conjunto de dados de teste

Para compreender melhor os erros cometidos pelo modelo de melhor desempenho, que utilizou o conjunto de dados *OC Score Ponderado Médio* em conjunto com o método de balanceamento *SMOTETomek*, NOGUEIRA et al. (2024) realizaram uma análise qualitativa dos *tweets* classificados incorretamente durante a fase de teste. Esta análise começa com a apresentação dos erros mais frequentes cometidos pelo modelo, conforme detalhado na Tabela 10.

Tabela 10 – Número total de *tweets* classificados incorretamente. **Fonte:** Traduzido de NOGUEIRA et al. (2024).

Tipo de erro	**Total de classificações incorretas**
Negativo classificado como neutro	126
Negativo classificado como positivo	165
Neutro classificado como negativo	707
Neutro classificado como positivo	668
Positivo classificado como negativo	214
Positivo classificado como neutro	276

Este resultado indica que os erros mais frequentes ocorreram quando os *posts* com sentimentos neutros foram erroneamente classificados como negativos ou positivos. Essas classificações incorretas podem ser atribuídas a diversos fatores, como a ambiguidade natural da linguagem, o uso de gírias e expressões coloquiais, além das interações específicas entre os usuários nas redes sociais, que frequentemente envolvem contextos informais e variáveis.

Para ilustrar como essas características podem afetar modelos de linguagem, NOGUEIRA et al. (2024) apresentaram três publicações do conjunto de dados de teste. O primeiro exemplo, descrito como: "Comi a pizza da Julia, comi a torta que minha cunhada comprou para meu sogro e agora vou comer sorvete com calda de chocolate que o Lucas trouxe para mim. Não quero guerra com ninguém", pode ser interpretado de diferentes maneiras, dependendo do contexto e das emoções subjacentes. Ao analisar cada segmento dessa frase, os autores observaram o seguinte:

- "Comi a pizza da Julia": Esta parte da frase é neutra, podendo até ser interpretada de forma positiva, pois não há indicações de conflito ou insatisfação. Ela pode transmitir uma sensação de prazer associada ao ato de comer a pizza.
- "Comi a torta que minha cunhada comprou para meu sogro". Esta parte da frase também pode ser classificada como neutra, uma vez que descreve uma ação simples, comer uma torta, sem evocar emoções intensas ou significativas.
- "Agora vou comer sorvete com calda de chocolate que o Lucas trouxe para mim". Novamente, esta parte pode ser interpretada como neutra ou positiva, pois descreve uma situação agradável, associada ao prazer de comer sorvete com calda de chocolate, especialmente quando essa ação é compartilhada com alguém que trouxe o doce.
- "Não quero guerra com ninguém.": Esta parte da frase altera o tom, sugerindo um sentimento negativo. No entanto, é importante destacar que a expressão é frequentemente utilizada de forma irônica ou figurativa, especialmente em redes sociais, e não necessariamente reflete um desejo genuíno de conflito ou hostilidade. Em muitos casos, ela transmite justamente o oposto: a ideia de que a pessoa está tranquila e em paz. Assim, embora a frase possa, à primeira vista, parecer indicar agressividade ou raiva, ela é usada, de forma contraditória, para expressar satisfação ou a ausência de conflito.

Em resumo, as primeiras partes da frase transmitem sentimentos neutros ou positivos, ligados ao prazer de comer pizza, torta e sorvete. No entanto, a última parte introduz uma ambiguidade que dificulta a análise do sentimento pelo modelo. Devido a essa incerteza, o modelo acaba classificando erroneamente a frase como se a mesma transmitisse um sentimento negativo.

Outro aspecto a ser considerado destacado por NOGUEIRA et al. (2024) é o contexto linguístico. Certos textos podem incluir referências culturais ou regionais que o modelo precisa interpretar adequadamente. No caso do OC, o contexto é relacionado ao tema de alimentos, sendo comum encontrar publicações que envolvem receitas culinárias, como ilustrado no próximo exemplo, que apresenta instruções para preparar um sanduíche:

- Corte o topo do pão.
- Pincele azeite de oliva por dentro do pão, pimenta-do-reino e um toque de sal.
- Adicione uma camada de queijo coalho ralado.
- Coloque uma porção de linguiça fatiada + presunto.
- Adicione 3 colheres de sopa de *cream cheese* e queijo *cheddar*.

Esta publicação descreveu as etapas para fazer sanduíches, com o objetivo principal de fornecer instruções para a preparação da receita. Embora ela tenha uma classificação neutra, o modelo pode interpretar erroneamente termos associados aos nomes dos alimentos como expressões positivas, resultando em uma classificação incorreta.

Outro exemplo relevante envolve o uso de gírias contemporâneas e postagens em redes sociais, onde os usuários frequentemente compartilham conteúdo com o objetivo de gerar alta interação. Essas postagens, no entanto, costumam incluir elementos que podem ser interpretados de forma ambígua pelos modelos de análise, resultando em classificações incorretas. Um exemplo ilustrativo é a seguinte postagem: “Coisas inúteis sobre mim: Altura: 1,81 metros, Idade: 15, Tamanho do pé: 42 ou 43, acho?, MBTI: Não lembro, Tatuagem: Nenhuma, *Piercing*: Nenhum, Cor favorita: Azul e preto, Anime favorito: *One Piece*, Animal favorito: Gato, Comida favorita: Lasanha, e para sobremesa, sorvete, Bebida favorita: Coca-Cola e Guaravita”. Nesse caso, a falta de contexto emocional explícito e o uso de linguagem informal podem levar a uma interpretação equivocada da polaridade do sentimento, destacando os desafios enfrentados pelos modelos ao analisar textos com características semelhantes.

A publicação, mais uma vez, faz referência a produtos e neste caso é o sorvete. Termos como esse podem ter influenciado a classificação da postagem, que foi inicialmente considerada neutra e informativa, para uma classificação positiva. Modelos de linguagem, assim como outros sistemas automatizados, enfrentam desafios significativos para captar as sutilezas da linguagem humana, especialmente em contextos informais e coloquiais. Essa limitação pode comprometer a precisão das classificações, gerando interpretações que nem sempre refletem o real teor da mensagem.

Com este estudo NOGUEIRA et al. (2024) demonstraram a viabilidade e a eficácia do uso de abordagens automatizadas para rotulagem de dados e construção de modelos de

análise de sentimentos aplicados ao contexto dos produtos lácteos no Brasil. A integração de cinco APIs distintas permitiu a criação de um conjunto robusto de 96.645 *tweets*, cuja rotulagem automática possibilitou a experimentação com diversas estratégias de combinação de resultados, evidenciando a variedade de metodologias disponíveis para a análise de sentimentos.

Um dos principais desafios enfrentados foi o desequilíbrio entre as classes negativa, neutra e positiva. A aplicação de técnicas de balanceamento de dados, como *SMOTE*, *ADASYN*, *SMOTE-ENN* e *SMOTETomek*, foi essencial para mitigar esse problema gerando conjuntos de dados mais equilibrados. A análise dos resultados mostrou que, embora algumas técnicas tenham apresentado desempenho superior durante o treinamento, nem sempre mantiveram essa performance na fase de teste, o que reforça a complexidade inerente à generalização dos modelos.

Dentre as abordagens avaliadas, os modelos balanceados com *SMOTETomek* destacaram-se, alcançando as melhores precisões médias nos conjuntos de dados *OC Label Dominante* e *OC Score Ponderado Médio*, com 77,17% e 79,61% de acurácia, respectivamente. O método *ADASYN* também demonstrou resultados promissores, especialmente no conjunto *OC Score Médio*, com precisão média de 75,45%. Esses resultados indicam que, apesar dos desafios relacionados ao *overfitting* e à generalização, as técnicas aplicadas foram eficazes para garantir um desempenho consistente em dados não vistos.

Em síntese, o estudo atingiu seu objetivo de desenvolver um modelo de análise de sentimentos eficiente e adaptado ao contexto específico dos *tweets* sobre produtos lácteos no Brasil. Os modelos resultantes demonstraram capacidade satisfatória de generalização, sendo adequados para aplicação no "Observatório do Consumidor". Este trabalho não apenas contribui para a análise automatizada de sentimentos em redes sociais, mas também abre caminho para futuras pesquisas que busquem aprimorar a precisão e a robustez desses modelos em contextos semelhantes, consolidando-se como uma base sólida para estudos posteriores na área.

Durante os experimentos, NOGUEIRA et al. (2024) buscaram desenvolver modelos de *deep learning*, o que resultou em um aumento expressivo dos custos computacionais. Esse impacto foi ainda mais acentuado pela combinação dos conjuntos de dados com diferentes técnicas de balanceamento. As abordagens de geração de amostras sintéticas, embora eficazes na mitigação do desequilíbrio entre as classes (negativa, neutra e positiva), também ampliaram significativamente o volume de dados. Em um dos testes, o modelo implementado exigiu mais de 213 GB de memória, ultrapassando a capacidade dos recursos disponíveis e inviabilizando sua execução. Essa limitação destacou os desafios computacionais enfrentados e representou um obstáculo relevante para a implementação deste modelo nesta pesquisa.

Outras limitações destacadas por NOGUEIRA et al. (2024) incluem a decisão de

aplicar a técnica de *Back Translation* exclusivamente à classe negativa do conjunto de dados *OC Label Dominante*, devido ao alto custo computacional associado. Além disso, os modelos enfrentaram desafios significativos na interpretação das complexidades da linguagem humana. Apesar dos avanços recentes em modelos de linguagem, eles ainda apresentam dificuldades para captar plenamente as nuances do idioma, especialmente em contextos informais e coloquiais, o que pode comprometer a precisão das classificações. Essas limitações ressaltam a necessidade contínua de aprimoramento dos modelos de análise de sentimentos.

5.2 Resultados de Alguns Estudos Desenvolvidos com as Análises Implementadas pelo OC

Com o uso da ferramenta do OC, diversos estudos e análises foram desenvolvidos ao longo da pesquisa. Cada trabalho buscou investigar os produtos lácteos de duas maneiras: de forma específica, analisando um produto em particular, e de forma ampla, explorando o segmento de produtos lácteos como um todo. Nesta seção serão apresentados os resultados de sete trabalhos (i) Mineração de dados em *tweets* para análise do consumo de lácteos no Brasil, (ii) Quais os derivados lácteos mais consumidos no Brasil durante a pandemia da COVID-19?, (iii) Análise exploratória do interesse por lácteos orgânicos, (iv) Impacto da inflação nos *tweets* sobre leite e derivados, (v) Observatório do Consumidor: potencializando a indústria láctea com *business intelligence* (vi) Observatório do Consumidor: explorando os consumidores de lácteos no Brasil, e (vii) Padrões de consumo de queijos no Brasil: Uma abordagem por meio do Observatório do Consumidor.

5.2.1 Trabalhos "Quais os derivados lácteos mais consumidos no Brasil durante a pandemia da COVID-19?" e "Mineração de dados em *tweets* para análise do consumo de lácteos no Brasil"

Os estudos (NOGUEIRA et al., 2022a) e (NOGUEIRA et al., 2022b) analisaram o consumo de lácteos por meio do Observatório do Consumidor, aplicando a análise comportamental em dois períodos distintos. Utilizando a mesma metodologia, ambos os trabalhos buscaram compreender o consumo dos lácteos explorando as relações entre ações (verbos) e produtos (substantivos) mencionados no Twitter.

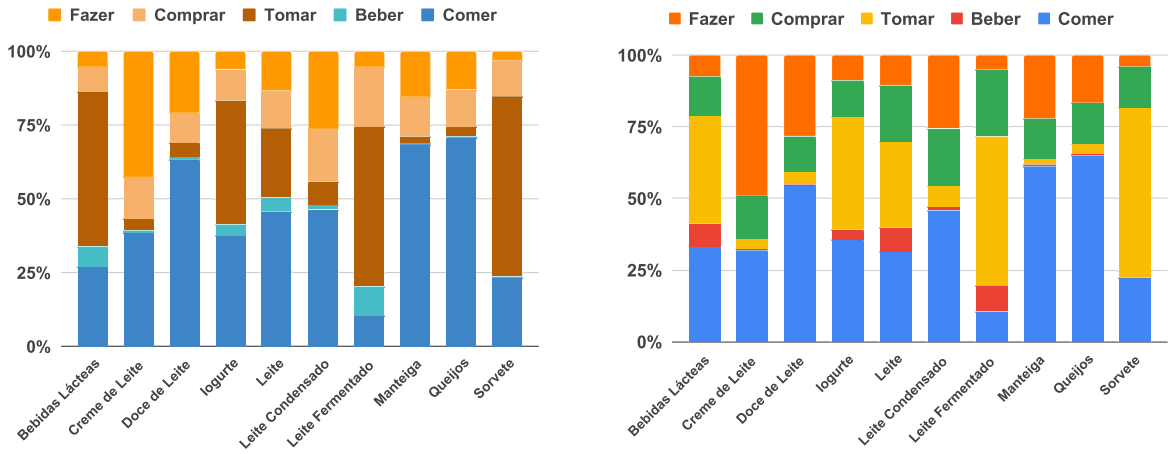
No trabalho de NOGUEIRA et al. (2022a) o foco foi em analisar os impactos da pandemia da COVID-19 no consumo de produtos lácteos no Brasil. Este estudo buscou responder à seguinte questão de pesquisa "Quais foram os derivados lácteos mais consumidos no Brasil durante a pandemia da COVID-19?" no período compreendido entre 07/05/2020[1] e 20/11/2021[2]. Já no estudo conduzido por NOGUEIRA et al. (2022b) o

[1] Início da coleta de dados do Observatório do Consumidor.
[2] Marca de 60% da população brasileira com o protocolo inicial de vacinação contra a COVID-19

objetivo foi o de analisar o contexto geral do consumo dos lácteos no Brasil para responder duas questões principais que foram: "Quais são os derivados lácteos mais consumidos no Brasil?" e "Como foi este consumo ao longo do tempo?", analisando-os por um período maior, que também iniciou no dia 07/05/2020 e foi até o dia 11/08/2022[3].

Durante a pandemia, NOGUEIRA et al. (2022a) observaram que a mudança abrupta na rotina impulsionou o consumo de conteúdos em redes sociais e de determinados alimentos. Nesse contexto, foram coletadas 19.033.823 publicações, das quais 70,3% mencionaram alguma ação relacionada ao consumo. No estudo posterior em que NOGUEIRA et al. (2022b) abrangeram um período mais longo, foram analisados 19.708.498 *tweets* sobre lácteos, com 70,4% contendo ao menos um verbo associado ao ato de consumir, direta ou indiretamente.

O foco dos estudos foi a extração de informações sobre o consumo, analisando especificamente a ocorrência de cinco verbos principais[4] e suas flexões verbais. Para o primeiro estudo, 1.705.383 *tweets* foram selecionados por serem publicados durante a pandemia e mencionarem verbos de interesse. No segundo estudo, que considerou um período mais amplo, a filtragem dos dados resultou em 2.410.272 publicações com verbos de consumo. A Figura 26 detalha a proporção de menções a cada verbo de consumo associado a esses produtos para cada um dos estudos.



(a) - Trabalho de NOGUEIRA et al. (2022a) (b) - Trabalho de NOGUEIRA et al. (2022b)

Figura 26 – Proporção de menções aos verbos de consumo por derivado lácteo. Em (a) resultado do trabalho que considera o período pandêmico. Em (b) resultado do trabalho que analisa um período mais amplo. **Fonte:** Retirado de NOGUEIRA et al. (2022a) e NOGUEIRA et al. (2022b).

Após contabilizar o número de menções aos verbos de consumo, os resultados mostraram que durante a pandemia, os produtos, sorvete (604.784), leite condensado

(primeira e segunda dose) concluída (MATHIEU et al., 2021).

[3] Último registro de dados coletado pelo Observatório do Consumidor neste estudo.

[4] Comer, beber, tomar, comprar e fazer.

(189.435), queijos (155.483), doce de leite (83.523) e a manteiga (61.107) foram os 5 lácteos mais consumidos. Por outro lado, ao analisar o período completo de dados há somente uma mudança. Os produtos, sorvete (851.917), leite condensado (541.972), queijos (257.311) e doce de leite (227.568) mantiveram a ordem vista no estudo anterior. A única diferença foi a aparição do leite (101.176) que obteve maior consumo que a manteiga neste período.

A análise dos resultados da Figura 26 revela padrões distintos de consumo para cada lácteo. Os verbos "fazer" e "comprar", ligados à aquisição de ingredientes para receitas, destacaram os produtos creme de leite, doce de leite, leite, leite condensado e manteiga. O verbo "tomar" teve mais menções para sorvete, leite fermentado, bebidas lácteas e iogurte, com o sorvete sendo o mais consumido, com 86,9% das menções durante a pandemia e 81,03% no período total. O verbo "beber" teve menos destaque, mas produtos lácteos líquidos como leite fermentado, bebidas lácteas, leite e iogurte foram mais mencionados.

O verbo "comer" esteve presente em todos os produtos analisados, com maior destaque para queijos, manteiga, doce de leite e leite condensado. Durante a pandemia, a proporção de menções aos verbos de consumo foi ligeiramente maior para quase todos os lácteos, destacando o consumo elevado de doce de leite, manteiga, leite condensado e queijos. Os queijos foram mais consumidos isoladamente ou em lanches *delivery*, enquanto a manteiga foi popular em receitas e no café da manhã. Os produtos indulgentes, como doce de leite e leite condensado, foram consumidos amplamente em diferentes momentos e situações.

Esses resultados corroboram os achados de SIQUEIRA et al. (2020a) e SIQUEIRA et al. (2020b). Os produtos lácteos mais consumidos, com exceção da manteiga, são caracterizados indulgentes e o pico de consumo destes produtos ocorreu no período compreendido entre maio e dezembro de 2020 como pode ser observado na Figura 27.

A Figura 27 mostra que o consumo de lácteos diminuiu ao longo dos meses, tendência confirmada pelo setor (RUGGIERO, 2022). Entre os cinco produtos em destaque, o sorvete foi o mais consumido, atingindo seu pico em outubro de 2020, com 58.357 menções aos verbos de consumo. O leite condensado registrou um pico em janeiro de 2021, impulsionado por um episódio amplamente debatido nas redes sociais: a divulgação do alto valor gasto pelo Governo Federal na compra do produto. Já o leite manteve um consumo relativamente estável, mas ganhou destaque a partir de setembro de 2021, quando novas palavras-chave foram adicionadas ao monitoramento do OC, ampliando a identificação de menções ao produto.

Os trabalhos demonstraram que a coleta, processamento e análise de dados de redes sociais pelo Observatório do Consumidor inovaram a abordagem tradicional de pesquisa de mercado. O avanço metodológico em relação ao trabalho de NOGUEIRA et al. (2021), que não analisava as conexões entre palavras nos *tweets*, possibilitou uma investigação mais aprofundada. Esse progresso permitiu, em trabalhos subsequentes, a exploração de

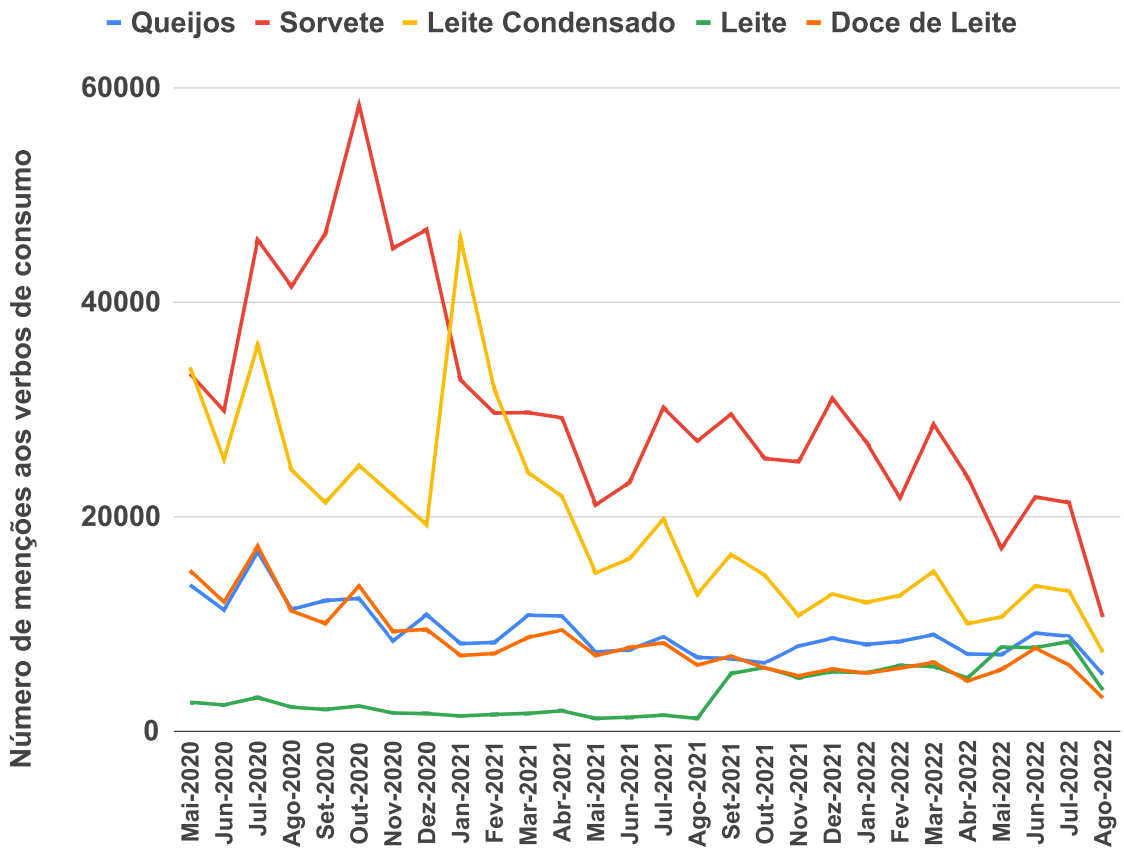


Figura 27 – Número de menções aos verbos de consumo ao longo do tempo dos 5 derivados lácteos mais representativos. **Fonte:** Retirado de NOGUEIRA et al. (2022b).

questões adicionais, como: “Qual o sentimento após o consumo?”, “Quem consumiu?”, “Quando consumiu?” e “Quais estados brasileiros apresentaram maior consumo?”, além da análise de dados de outras redes sociais.

5.2.2 Trabalho “Análise exploratória do interesse por lácteos orgânicos”

O estudo de SILVA et al. (2022) investigou o interesse por lácteos orgânicos no Brasil, um tema crescente no debate sobre produtos alimentícios. A disputa entre produtos orgânicos e não orgânicos tem ganhado destaque, e o mercado lácteo não é exceção. Apesar da variedade de produtos lácteos convencionais, o surgimento dos orgânicos aumentou as opções para o consumidor. No entanto, pouco se sabe sobre o tamanho desse mercado no Brasil, especialmente devido às dificuldades e custos das pesquisas tradicionais em um país de grandes dimensões.

O estudo adaptou o módulo de coleta de dados do Observatório do Consumidor para analisar o interesse por lácteos orgânicos no Twitter, focando em produtos orgânicos e não orgânicos nos países Brasil, Alemanha, Estados Unidos, Espanha, Itália, França e Holanda. As palavras-chave foram traduzidas para os idiomas locais e a região geográfica

de cada país foi incorporada ao *script* para garantir a precisão dos dados. A coleta ocorreu entre 28 de junho e 05 de julho de 2022, totalizando 799.888 postagens, comparando menções a produtos orgânicos e não orgânicos, conforme apresentado na Tabela 11.

Tabela 11 – Quantidade de *tweets* sobre lácteos (orgânicos ou não) por país. **Fonte:** Retirado de SILVA et al. (2022).

País	**Quantidade de *tweets***		
	Orgânicos	**Não-Orgânicos**	**Total**
Alemanha	4	13.682	**13.686**
Brasil	0	84.988	**84.988**
Espanha	5	111.661	**111.666**
Estados Unidos	112	522.461	**522.583**
França	10	9.845	**9.855**
Holanda	17	5.256	**5.273**
Itália	0	31.847	**31.847**
Total	**148**	**799.740**	**799.888**

Na Tabela 11, observa-se que os Estados Unidos lideraram em número absoluto de *tweets* sobre derivados lácteos orgânicos, com 0,021% dos *tweets* americanos (112 de 522.583) focando nesses produtos, alinhando-se com o fato de os EUA serem o maior mercado de lácteos orgânicos, conforme IFOAM GENERAL ASSEMBLY (2008).

No entanto, a Holanda, com 0,325% das postagens sobre lácteos dedicadas aos orgânicos, é o país com maior interesse proporcional. Isso corresponde a 17 *tweets* entre os 5.273 sobre derivados lácteos no período. A França ocupa o segundo lugar, com 0,101% das postagens focadas em lácteos orgânicos. De acordo com IFOAM GENERAL ASSEMBLY (2008), a França é o terceiro maior mercado de alimentos orgânicos, enquanto a Holanda ocupa a 13ª posição. Segundo KPMG (2018), a França representa 7,7% do mercado global de leite orgânico e a Holanda 0,3%. Brasil e Itália não apresentaram *tweets* sobre lácteos orgânicos, indicando um interesse ainda incipiente nesses mercados.

Este trabalho destaca a versatilidade do Observatório do Consumidor, que, embora focado no universo lácteo, pode ser facilmente adaptado para outros contextos com ajustes nas palavras-chave. A pesquisa sobre lácteos orgânicos demonstrou que, embora o interesse dos usuários do Twitter seja ainda modesto em comparação com derivados lácteos em geral, a ferramenta se mostrou eficaz na mensuração desse mercado em expansão.

No Brasil, o interesse por lácteos orgânicos é limitado, possivelmente devido a fatores econômicos. A pesquisa evidencia a utilidade do Observatório do Consumidor para captar as dinâmicas desse mercado. Futuros estudos podem explorar outras redes sociais para ampliar a compreensão sobre o consumo de lácteos orgânicos.

5.2.3 Trabalho “Impacto da inflação nos *tweets* sobre leite e derivados”

O estudo de FRANCO et al. (2022) investigou a correlação entre a inflação e o sentimento negativo dos consumidores de leite e derivados, associando a quantidade de *tweets* sobre esses produtos com a inflação. A principal hipótese do trabalho era que variáveis econômicas, como desemprego, renda e inflação, influenciavam a experiência do consumidor e impactavam a análise de sentimentos das publicações coletadas pelo OC.

FRANCO et al. (2022) destacam que, sob a condição de *ceteris paribus*, a inflação — especialmente quando acompanhada pela estagnação da renda — reduz o poder de compra dos indivíduos, gerando um sentimento negativo. Com base nessa premissa, o estudo buscou analisar como a percepção dos consumidores de produtos lácteos variava em diferentes cenários econômicos, particularmente em períodos de alta inflação.

Após realizar a análise de sentimentos nas publicações, selecionaram-se aquelas classificadas com sentimento negativo. Para analisar a inflação, foram utilizados os dados do Índice de Preços ao Consumidor Amplo (IPCA) do IBGE/SIDRA (2022), com o período de análise compreendendo de maio de 2020 a junho de 2022.

A fim de avaliar o impacto da inflação sobre o sentimento expresso nos *tweets* sobre leite e derivados, foi adotada a variação na porcentagem de postagens com conteúdo negativo, conforme ilustrado na Figura 28.

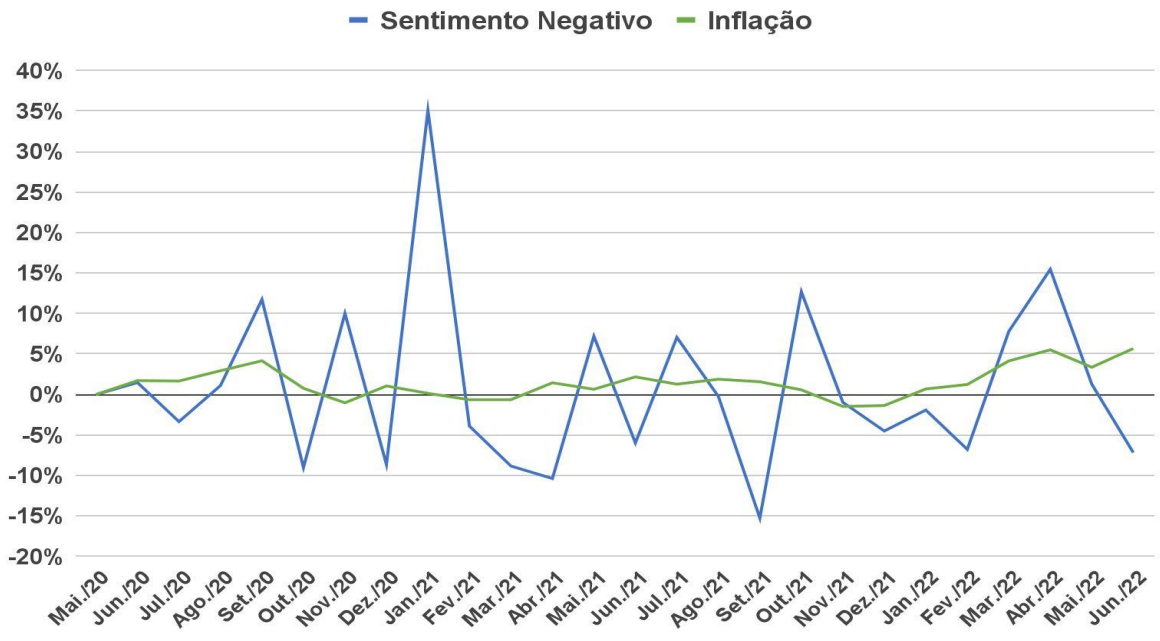


Figura 28 – Variação da porcentagem de *tweets* com sentimento negativo e a inflação. **Fonte:** Retirado de FRANCO et al. (2022).

A Figura 28 mostra que não há uma relação direta entre a variação de preços e o sentimento negativo nos *tweets* sobre leite e derivados. Deste modo, não é possível afirmar que a inflação impacta o sentimento negativo do consumidor. Apesar de o preço influenciar

o consumo de lácteos, o estudo indica que ele não afeta o sentimento expresso. Isso evidencia que a análise de sentimentos reflete, sobretudo, as preferências dos consumidores.

Este trabalho demonstrou que os resultados das análises realizadas pelo Observatório do Consumidor permitem correlacionar dados com variáveis externas, como o IPCA. Embora não tenha sido identificada uma relação direta entre a variação dos preços e o sentimento expresso nos *tweets* sobre lácteos, ficou claro que, embora a inflação seja uma variável importante no consumo de lácteos, ela não afetou o sentimento do consumidor em relação aos produtos.

5.2.4 Trabalho "Observatório do Consumidor: potencializando a indústria láctea com *business intelligence*"

Neste estudo, NOGUEIRA et al. (2024) concentraram-se na análise de uma categoria específica de lácteos — os queijos — com o objetivo de compreender o perfil dos consumidores desse produto no Brasil. Para isso, responderam às seis perguntas-chave propostas pelo OC, que serviram de base para a construção de uma persona representativa desses consumidores: (i) quem são os consumidores de queijo?; (ii) quando ocorre o consumo desses produtos?; (iii) qual é o sentimento associado a esse consumo?; (iv) quais tipos de queijos se destacam?; (v) quais são os principais acompanhamentos?; e (vi) em quais regiões o queijo é mais consumido?

Além da rede social X, os dados do YouTube também foram incluídos nas análises. Especificamente em relação aos dados sobre queijos, foco deste estudo de caso, foram identificadas 8.711.860 publicações, representando 30,6% do total de dados relacionados aos lácteos analisados pelo OC (28.465.878 publicações). Ao selecionar apenas as publicações em que haviam os verbos associados ao consumo ligados aos nomes dos queijos (substantivos) o universo de dados para todas as análises subsequentes consistiu em 230.939 postagens.

A análise realizada por NOGUEIRA et al. (2024) revelou que 74,7% das postagens sobre o consumo de queijos foram realizadas por homens, enquanto 25,3% foram feitas por mulheres. Entre os homens, 71,3% pertencem à geração Y, nascidos entre 1981 e 1995, e, entre as mulheres, essa porcentagem sobe para 77,2%. Em relação à renda, a classe D predominou entre os consumidores de ambos os sexos, representando 41,7% do total.

Para analisar a evolução do consumo no tempo, NOGUEIRA et al. (2024) destacam que o OC realizou uma análise temporal das menções a verbos relacionados ao consumo de alimentos. A Figura 29 ilustra a evolução do consumo ao longo do tempo.

De forma geral, NOGUEIRA et al. (2024) observaram picos no número de publicações sobre queijos no mês de julho, sugerindo um aumento no consumo durante esse período. O maior volume de postagens foi registrado durante a pandemia, quando as medidas de isolamento social mantiveram as pessoas em casa, possivelmente influenciando

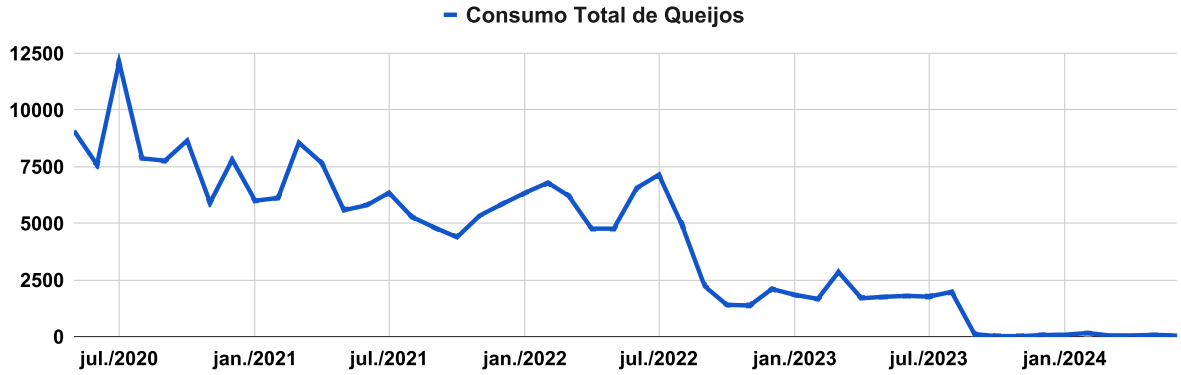


Figura 29 – Resultado da análise temporal das publicações sobre o consumo de queijos. **Fonte:** Retirado de NOGUEIRA et al. (2024).

seus hábitos alimentares. No entanto, é importante destacar que, apesar da redução nas postagens sobre queijos após esse período, mudanças nas políticas de compartilhamento de dados da rede social X podem ter impactado a coleta de informações pelo OC, o que pode ter influenciado a análise temporal e a interpretação dos resultados.

No que se refere à avaliação do consumo, este estudo adotou a análise de sentimentos proposta por NOGUEIRA et al. (2024) para classificar as publicações relacionadas ao consumo em três categorias: negativas, neutras ou positivas. Os resultados obtidos demonstraram que 78,5% das publicações pós-consumo expressaram sentimentos positivos, evidenciando um alto nível de satisfação dos consumidores em relação ao queijo. Esse achado reforça a percepção favorável do produto no mercado e sugere uma forte aceitação por parte do público.

Ao analisar os tipos de queijos mais mencionados, NOGUEIRA et al. (2024) constataram que o queijo coalho foi o mais citado, com 6.125 menções no período estudado. Esse tipo de queijo é amplamente popular em diversas regiões do Brasil, especialmente por sua versatilidade e sabor único, sendo frequentemente utilizado em churrascos e pratos típicos da culinária nordestina. Em segundo lugar, o requeijão aparece com 4.694 menções, destacando-se por sua textura cremosa e uso comum em pães, torradas e uma variedade de receitas. O queijo manteiga e o queijo minas ocupam as posições seguintes, com 2.776 e 2.631 menções, respectivamente, sendo ambos valorizados por sua qualidade e ampla aplicação na gastronomia brasileira. Outros tipos, como queijo ralado, mussarela, prato, *cheddar*, sem lactose e canastra, também foram mencionados, evidenciando a diversidade de preferências e necessidades dos consumidores, que buscam desde opções tradicionais até alternativas adaptadas a restrições dietéticas. Essa variedade reflete a riqueza da cultura queijeira no Brasil e a importância desses produtos no cotidiano alimentar da população.

Para identificar os principais acompanhamentos mencionados nas publicações relacionadas aos queijos, a Figura 30 apresenta uma *word cloud* detalhada, evidenciando os termos mais frequentemente associados. Essa visualização destaca, de forma intuitiva,

os acompanhamentos que aparecem com maior recorrência, permitindo uma compreensão rápida das preferências e combinações mais populares relatadas pelos usuários.

Figura 30 – *Word cloud* dos principais acompanhamentos dos queijos. Obs: em azul-claro, produtos categorizados como in natura e/ou minimamente processados; em azul-escuro, produtos ultraprocessados; em laranja, produtos processados; e em amarelo, ingredientes culinários, segundo o Guia Alimentar para a População Brasileira (MINISTÉRIO DA SAÚDE, 2014). **Fonte:** Retirado de NOGUEIRA et al. (2024).

Ao analisar a *word cloud*, NOGUEIRA et al. (2024) observaram que, embora os acompanhamentos *in natura* e minimamente processados fossem os mais frequentes, havia também uma presença expressiva de alimentos processados e ultraprocessados. Esse resultado reflete a complexidade da alimentação contemporânea, em que a busca por praticidade e conveniência muitas vezes se sobrepõe à preferência por opções mais nutritivas. A análise evidencia padrões de consumo que destacam informações sobre os hábitos alimentares da população, que podem orientar estratégias mais assertivas para o setor alimentício, tanto no desenvolvimento de produtos quanto em campanhas de *marketing* e educação nutricional.

A análise da distribuição regional das publicações sobre o consumo de queijo revelou que a região Sudeste concentrou a maioria das menções, representando aproximadamente 53,13% do total. Em seguida, destacou-se a região Nordeste, com cerca de 21,33%, enquanto o Sul respondeu por aproximadamente 15,04% das postagens. A região Centro-Oeste contribuiu com 7,70% das publicações, e a região Norte apresentou a menor participação, com 3,92%. É importante ressaltar que Pernambuco, localizado no Nordeste, figura entre os estados com maior consumo de queijos, evidenciando uma forte tradição no consumo de queijos artesanais. Entre eles, o queijo coalho destacou-se como o mais citado nas

publicações, refletindo sua popularidade e relevância cultural na região.

Como conclusões, NOGUEIRA et al. (2024) ressaltam que os dados analisados revelaram tendências de mercado significativas, evidenciando a importância de estratégias regionais diferenciadas. No Sudeste, onde o consumo é mais consolidado, campanhas de fidelização e o lançamento de novos produtos podem fortalecer ainda mais a presença das marcas. Em contrapartida, o Norte representa um mercado com grande potencial de expansão, exigindo iniciativas que ampliem a penetração e a acessibilidade dos queijos na região. Essa abordagem regional é fundamental para orientar decisões estratégicas das empresas do setor, otimizando ações de *marketing* e distribuição para maximizar o alcance e a eficácia das campanhas.

Além disso, os resultados obtidos pelo OC proporcionam uma visão abrangente e atualizada sobre o comportamento dos consumidores de queijos no Brasil. A análise das publicações indicou que a maioria dos comentários após o consumo possui um tom positivo, refletindo altos níveis de satisfação com os produtos. Com o suporte do OC, as empresas do setor não apenas ganham a capacidade de monitorar e entender as tendências de consumo, mas também de se adaptar rapidamente às mudanças nas demandas do mercado. Essa agilidade na tomada de decisões é crucial para impulsionar o crescimento sustentável e fortalecer a competitividade no mercado de queijos no Brasil, garantindo que as ações implementadas estejam alinhadas com as expectativas e necessidades dos consumidores.

5.2.5 Trabalho “Observatório do Consumidor: explorando os consumidores de lácteos no Brasil”

Em outro estudo, OLIVEIRA et al. (2024) concentraram-se na análise das características do perfil dos consumidores de lácteos no Brasil. A pesquisa foi orientada pela seguinte questão norteadora: “Quem são os consumidores de produtos lácteos?”. Para responder a essa pergunta, os autores utilizaram o OC como ferramenta principal, explorando dados da rede social X para identificar padrões de comportamento, preferências e características demográficas associadas ao consumo de produtos lácteos no país.

Ao todo, OLIVEIRA et al. (2024) selecionaram 640.890 publicações a partir de um universo de mais de 23 milhões de postagens coletadas. A seleção considerou publicações que continham verbos associados ao consumo de produtos lácteos, conforme detalhado na subseção 4.1.4.7. Posteriormente, os dados foram ainda mais refinados, restringindo-se às postagens cujas fotos de perfil permitiam a extração de informações demográficas, como sexo e idade dos usuários. Esse processo resultou em um conjunto final de 280.268 publicações, oferecendo uma base mais precisa para a análise do perfil dos consumidores.

A análise conduzida por OLIVEIRA et al. (2024) revelou que o consumo de produtos lácteos foi predominantemente masculino, com 73,53% das publicações atribuídas a homens

e 26,47% a mulheres. Dentre os consumidores do sexo masculino, 52,11% pertencem à Geração Y, nascidos entre 1981 e 1995, destacando a relevância desse grupo etário no mercado de lácteos. No que diz respeito à renda, a classe D, composta por indivíduos com rendimentos entre 2 e 3 salários mínimos, foi a mais representativa entre os usuários, independentemente do gênero. As profissões mais mencionadas nas publicações incluíram jornalistas, advogados e nutricionistas, sugerindo uma diversidade de perfis profissionais entre os consumidores de lácteos.

As informações obtidas por OLIVEIRA et al. (2024) são de grande relevância para as empresas do setor lácteo, pois proporcionam uma compreensão aprofundada do perfil e comportamento dos consumidores brasileiros. Essas informações permitem a elaboração de estratégias de *marketing* mais precisas e segmentadas, alinhadas às necessidades e preferências do público-alvo. Com o suporte do OC, as organizações podem tomar decisões mais fundamentadas, potencializando suas chances de sucesso e fortalecendo sua competitividade no dinâmico mercado atual.

5.2.6 Trabalho "Padrões de consumo de queijos no Brasil: Uma abordagem por meio do Observatório do Consumidor"

Os estudos conduzidos pelo OC destacaram amplamente a análise de queijos, tema impulsionado pelo grande volume de publicações coletadas durante a pesquisa e pela significativa importância econômica desse produto, derivada de seu alto consumo no Brasil. Conforme apontado por RODRIGUES et al. (2024), o consumo de queijos corresponde a 7% do total de lácteos consumidos no país (IBGE, 2020). Esse dado evidencia a importância do queijo na alimentação dos brasileiros, refletindo não apenas suas preferências gastronômicas, mas também elementos culturais e socioeconômicos profundamente enraizados na sociedade (RODRIGUES et al., 2024).

De acordo com RODRIGUES et al. (2024), a base de dados do IBGE (2020) estava desatualizada, com informações generalizadas e com uma defasagem de quase quatro anos. Para superar essa limitação, o estudo recorreu aos dados coletados pela ferramenta OC, proveniente da rede social X, que oferecia uma base constantemente atualizada. O foco da pesquisa foi analisar os padrões de consumo de queijos populares, finos e artesanais, proporcionando uma visão detalhada sobre os hábitos alimentares relacionados a esses produtos no Brasil.

Ao todo, foram analisadas 1.772.609 publicações sobre queijos na rede social X, no Brasil, durante o período de maio de 2020 a agosto de 2023. Dentre as postagens coletadas, a maioria se referia aos queijos populares (69%), seguida pelos queijos artesanais (26%) e queijos finos (5%). A análise de sentimentos, usando o modelo implementado por NOGUEIRA et al. (2024), revelou que, independentemente da categoria, as publicações sobre queijos foram predominantemente positivas: cerca de 80% das postagens sobre queijos

finos, 79% sobre queijos populares e 74% sobre queijos artesanais expressaram sentimentos favoráveis. Esses resultados indicaram que os consumidores brasileiros demonstraram uma forte apreciação pelos queijos, mostrando um interesse crescente por diferentes tipos de queijos, seja pela tradição, pelo sabor ou pela exclusividade das opções disponíveis.

Além disso, RODRIGUES et al. (2024), por meio de análises realizadas utilizando o OC como ferramenta, constataram que o perfil dos autores das postagens não apresentou variações significativas entre as diferentes categorias de queijos. De modo geral, as publicações foram predominantemente feitas por homens da Geração Y (com idades entre 25 e 40 anos), pertencentes à classe social D, com renda mensal de 1 a 3 salários mínimos. Geograficamente, a maior concentração de postagens foi observada na região Sudeste do Brasil, o que reflete o perfil demográfico e econômico do público mais ativo e engajado com o tema na rede social X.

Os dados analisados revelaram associações claras entre os tipos de queijos e seus acompanhamentos preferidos, que variaram conforme a categoria do produto, conforme ilustrado na Tabela 12.

Tabela 12 – *Ranking* dos tipos de queijos e seus acompanhamentos mais citados no X no Brasil **Fonte:** Adaptado de RODRIGUES et al. (2024).

Categoria de Queijo	**Tipo de Queijo**	**Acompanhamento**
Queijos Populares	Requeijão	Pão
	Danoninho	Pão de queijo
	Queijo Minas	Miojo
Queijos Finos	Queijo *Brie*	Pão
	Parmesão	Pimenta
	Gorgonzola	Tomate
Queijos Artesanais	Queijo Coalho	Pão
	Queijinho	Manteiga
	Queijo Manteiga	Leite

Os padrões de consumo observados destacam preferências distintas entre as categorias de queijos, refletindo não apenas aspectos culturais, mas também hábitos alimentares específicos dos consumidores no Brasil. Segundo RODRIGUES et al. (2024), o queijo mais mencionado nas postagens foi o requeijão, que representou 40% das menções a queijos na rede X. Em seguida, destacou-se o queijo coalho, com 8%, evidenciando a popularidade do requeijão, cuja versatilidade pode explicar seu alto índice de citações.

Entre os queijos finos, o *Brie* se destacou, representando 26% das menções dentro dessa categoria, embora corresponda a apenas 1% do total de menções a queijos em geral. RODRIGUES et al. (2024) ressaltam que, no Brasil, os queijos populares são os que despertam maior interesse entre os consumidores, consolidando-se como os mais citados e apreciados. Os queijos artesanais ocupam a segunda posição no *ranking* de popularidade, refletindo uma crescente valorização de produtos com características regionais e métodos

tradicionais de produção. Por fim, os queijos finos, embora presentes nas discussões, continuam a ser associados a um consumo mais elitizado, o que explica sua menor representatividade em comparação às outras categorias. Essa hierarquia de preferências evidencia as nuances do mercado de queijos no Brasil, marcado pela diversidade de perfis de consumo e pela influência de fatores culturais e socioeconômicos.

Pela análise dos acompanhamentos citados na Tabela 12, observa-se que os queijos populares são comumente associados a produtos práticos e típicos do dia a dia de brasileiros, indicando uma preferência por conveniência no momento do consumo. Já os queijos finos são vinculados a combinações mais sofisticadas, que realçam o sabor do produto, como pimentas e tomate, indicando uma inclinação ao consumo em momentos especiais. Os principais queijos citados na categoria foram o queijo *Brie*, Parmesão e Gorgonzola, caracterizados pelo sabor forte e característico, de forma a serem consumidos esporadicamente, em ocasiões particulares (PINTO et al., 2012).

Por fim, RODRIGUES et al. (2024) destacam que os queijos artesanais estiveram associados a produtos tradicionais e de base simples, como leite e manteiga, podendo ser utilizados tanto como ingredientes quanto como acompanhamentos. Essa característica reforça a versatilidade desses produtos, exemplificada pelo queijo Coalho, que foi consumido tanto como prato principal quanto como complemento em refeições. Além disso, o pão emergiu como o acompanhamento mais citado em todas as categorias, representando 29% das menções relacionadas a queijos. Esse dado indica que, independentemente do tipo de queijo, o pão foi o acompanhamento preferido e mais amplamente associado ao consumo de queijo no Brasil.

Apesar das diferenças nas preferências de acompanhamentos, RODRIGUES et al. (2024) mencionaram que o perfil do consumidor interessado em queijos mostrou-se consistente entre as três categorias analisadas, sendo predominantemente representado por homens da Geração Y, de classe média baixa.

Os resultados obtidos por RODRIGUES et al. (2024) evidenciaram distinções significativas nos padrões de consumo entre as diferentes categorias de queijo, apontando para motivações distintas por trás do interesse pelo produto. Os queijos populares mostraram-se mais presentes no consumo diário, refletindo sua acessibilidade e conveniência. Em contrapartida, os queijos finos foram associados a ocasiões mais elaboradas, como festas e confraternizações, sugerindo um consumo mais pontual e voltado para momentos especiais. Já os queijos artesanais destacaram-se por sua versatilidade, sendo utilizados tanto em preparações culinárias quanto em eventos especiais, o que reforça sua conexão com tradições e práticas gastronômicas regionais.

Com a utilização do OC foi possível extrair informações essenciais que podem direcionar estratégias de *marketing* e desenvolvimento de produtos, permitindo que as empresas atendam de maneira mais eficaz às necessidades e expectativas dos consumidores

brasileiros. Ao compreender as preferências e os contextos de consumo associados a cada categoria de queijo, é possível criar campanhas e produtos que ressoem com os diferentes perfis de consumidores, potencializando a satisfação e a fidelização do público.

5.3 Síntese das *Newsletter* Publicadas

Uma das funcionalidades oferecidas pela plataforma do OC são as *newsletters*. Ao todo, 29 edições foram publicadas no sistema, cujos resultados, em sua maioria, foram amplamente divulgados em veículos de informação especializados no setor. Essa estratégia ampliou significativamente o alcance da plataforma e reforçou sua relevância no mercado. A Tabela 13 apresenta o período e os temas das *newsletters* desenvolvidas pela equipe do OC.

Tabela 13 – Temas das *Newsletters* do Observatório do Consumidor (2022-2025)

Período	**Tema da *Newsletter***
Março - 2022	O que é o Observatório do Consumidor?
Abril - 2022	Funcionalidades do Observatório do Consumidor
Maio - 2022	Especial sobre iogurte
Junho - 2022	O interesse por lácteos no primeiro semestre de 2022
Julho - 2022	O país do leite condensado
Agosto - 2022	A questão do preço dos lácteos
Setembro - 2022	Especial sobre queijos
Outubro - 2022	O interesse por lácteos na Região Sul do Brasil
Novembro - 2022	O queijo como o centro das atenções
Dezembro - 2022	O leite no País do futebol
Janeiro - 2023	Os lácteos no verão: Uma análise do primeiro trimestre de 2021 e 2022
Fevereiro - 2023	Críticas aos queijos
Março - 2023	O consumo de doce de leite no Brasil
Abril - 2023	A Quaresma e a demanda por lácteos
Maio - 2023	O consumo de bebidas lácteas no Brasil
Junho - 2023	A transição nutricional dos lácteos no Brasil
Julho - 2023	Dinâmica populacional e lácteos
Agosto - 2023	O consumo de leite fermentado no pós-pandemia
Setembro - 2023	O consumo de queijos no último ano
Outubro - 2023	O interesse dos consumidores por lácteos orgânicos
Novembro - 2023	Sorvete ou doce de leite?
Dezembro - 2023	O consumo de lácteos e as mudanças climáticas
Fevereiro - 2024	*Milk shaming* (vergonha do leite)
Abril - 2024	Leite A2A2 - Percepção dos consumidores
Junho - 2024	O interesse por leite sem lactose
Agosto - 2024	O consumo do leite desnatado no Brasil
Outubro - 2024	*Insights* do YouTube sobre leite e derivados
Dezembro - 2024	Tendência de consumo de lácteos na temporada de final de ano
Fevereiro - 2025	Bebidas lácteas proteicas: tendência ou realidade consolidada?

Os temas abordados nas *newsletters* abrangem uma ampla gama de tópicos relacionados ao consumo de lácteos no Brasil. Essa diversidade possibilitou a realização de

análises detalhadas para compreender tendências de consumo, preferências regionais e fatores sociais e econômicos que influenciam o mercado.

Com base nesses estudos, oito tópicos principais foram explorados. O primeiro diz respeito à análise de tendências de consumo ao longo do tempo, com o objetivo de identificar mudanças no interesse por diferentes tipos de lácteos, como iogurtes, queijos e leite condensado, em períodos específicos. Além disso, foi possível avaliar o impacto de eventos sazonais e datas comemorativas no consumo, como nos estudos "A Quaresma e a demanda por lácteos" (abril de 2023) e "Tendência de consumo de lácteos na temporada de final de ano" (dezembro de 2024).

Outro tópico explorado foi a análise regional e sociocultural, investigando como as preferências locais influenciam o consumo. Exemplos incluem "O interesse por lácteos na Região Sul do Brasil" (outubro de 2022) e o impacto de práticas culturais, como a tradição de consumo de doce de leite, destacada em "Doce de leite: um patrimônio cultural e seu mercado" (março de 2023).

Outro tópico abordado foi o comportamento do consumidor, permitindo examinar percepções e sentimentos associados a novos produtos, como "Leite A2A2" (abril de 2024) e "O interesse por leite sem lactose" (junho de 2024). Além disso, foram avaliadas mudanças comportamentais, como o fenômeno do "*Milk Shaming*" (vergonha do leite), analisado em fevereiro de 2024, que reflete um maior foco em questões ambientais e éticas.

O impacto econômico também foi um aspecto explorado. Questões relacionadas ao custo foram analisadas, como apresentado em "A questão do preço dos lácteos" (agosto de 2022), para avaliar como mudanças econômicas influenciam o consumo de lácteos.

Outro eixo de análise abordou as preferências por produtos específicos, investigando tipos de produtos destacados pelos consumidores. Exemplos incluem "O país do leite condensado" (julho de 2022), "O consumo de bebidas lácteas no Brasil" (maio de 2023), "O interesse dos consumidores por lácteos orgânicos" (outubro de 2023) e "Bebidas lácteas proteicas: tendência ou realidade consolidada?" (fevereiro de 2025).

Questões de saúde e sustentabilidade também foram objeto de estudo. A relação entre o consumo de lácteos e a saúde foi analisada em "A transição nutricional dos lácteos no Brasil" (junho de 2023) e em "O interesse por leite sem lactose" (junho de 2024). Além disso, o impacto ambiental foi explorado em "O consumo de lácteos e as mudanças climáticas" (dezembro de 2023), avaliando como questões climáticas influenciam o comportamento dos consumidores.

A incorporação de outras redes sociais ao Observatório do Consumidor permitiu expandir as análises, como demonstrado no estudo "*Insights* do YouTube sobre leite e derivados" (outubro de 2024), que investigou como plataformas digitais impactam o consumo.

Por fim, a ferramenta também possibilitou análises comparativas entre diferentes produtos. Um exemplo é “Sorvete ou doce de leite?” (novembro de 2023), que buscou entender as preferências e comportamentos dos consumidores em relação a produtos concorrentes.

Essas análises não apenas ajudam a compreender as dinâmicas do mercado de lácteos, mas também oferecem subsídios para traçar perfis detalhados de consumidores e prever tendências futuras. A base de dados fornecida pelas *newsletters* constitui uma ferramenta rica e diversificada, com aplicações acadêmicas, mercadológicas e institucionais.

5.4 Resultados do Sistema Computacional

Diante de todo o contexto apresentado, nesta seção, apresenta-se como resultado algumas páginas do sistema desenvolvido. A Figura 31 mostra em detalhes a página inicial do sistema.

Figura 31 – Página inicial do sistema computacional do Observatório do Consumidor. **Fonte:** Observatório do Consumidor.

A primeira imagem do sistema encapsula a essência do OC, destacando qual é o propósito da ferramenta. Na parte superior direita da interface, são apresentadas duas opções projetadas para promover a inclusão e atender a algumas necessidades dos usuários: uma funcionalidade que permite aumentar ou diminuir o tamanho da fonte dos textos e uma opção de tradução para Libras (Língua Brasileira de Sinais). Essas funcionalidades visam tornar o sistema mais acessível, garantindo que diversas pessoas, independentemente de suas habilidades visuais ou auditivas, possam navegar e interagir com a plataforma de maneira eficaz.

Um pouco mais abaixo, é possível visualizar as diferentes páginas que o sistema oferece. A página “Início” permite ao usuário retornar à página inicial. Em seguida, encontra-se a página de “Análises”, em que o usuário pode escolher a plataforma de sua

preferência para realizar suas análises. A página de "*Newsletter*" apresenta todas as análises realizadas pela equipe do OC sobre diversos temas. Também está disponível a página "Sobre o Projeto", que fornece informações detalhadas sobre os produtos lácteos analisados, além de um glossário que explica diversos termos que podem ser novos para o usuário. Por fim, a página Institucional oferece uma seção "Saiba Mais", que resume informações sobre o projeto, fornecendo visualizações sobre a equipe do OC e os trabalhos publicados.

Ao clicar na página de análises que possui a parte mais central do sistema — as visualizações gráficas das análises — os usuários do OC precisam definir inicialmente o tipo de busca desejada. As opções de filtragem incluem critérios como regiões, tipos de produtos e sentimentos. Após selecionar os filtros, o próximo passo é especificar o período de análise e clicar no botão de pesquisa. Com essa etapa concluída, a página com os resultados da busca será carregada. A Figura 32 ilustra essa funcionalidade, destacando a página de busca de resultados, que, neste exemplo, foca na rede social X.

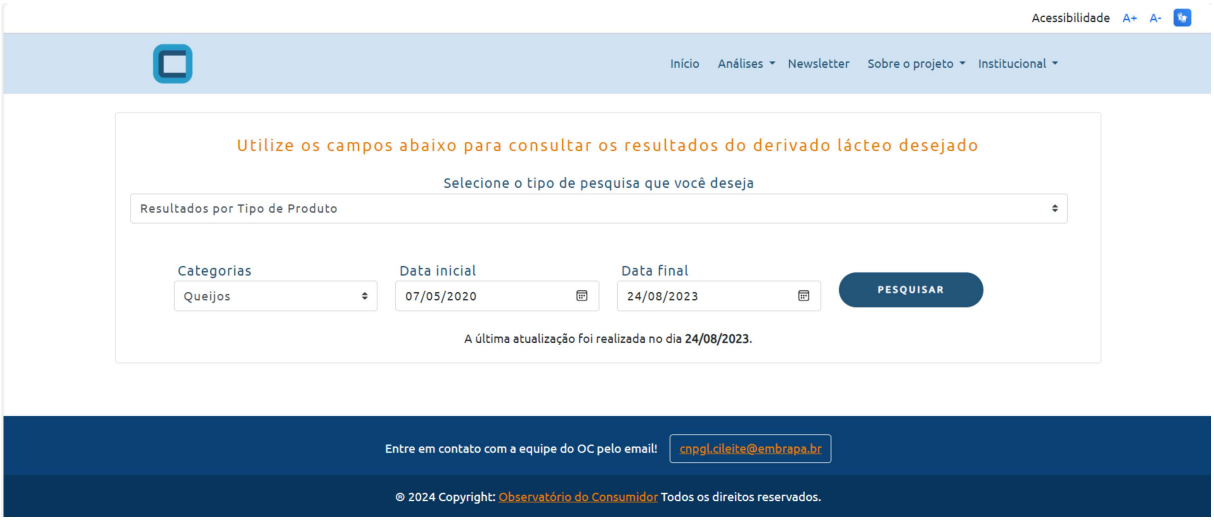


Figura 32 – Página de busca do sistema do Observatório do Consumidor. **Fonte:** Observatório do Consumidor.

Uma vez que o usuário tenha realizado essas etapas, os resultados são consultados, processados e exibidos na página, gerando as visualizações que serão exibidas iniciando com a Figura 33.

Inicialmente, são apresentados dois *cards*: um que exibe o total de postagens relacionadas ao interesse de consumidores e usuários da rede social X sobre queijos, e outro que mostra o número total de usuários envolvidos nessas discussões. Observa-se um volume expressivo de publicações, ultrapassando 8 milhões de menções, o que evidencia a popularidade e o amplo consumo desse produto entre os brasileiros. Além disso, nota-se um número significativo de usuários distintos que mencionaram o produto, somando mais de 6 milhões.

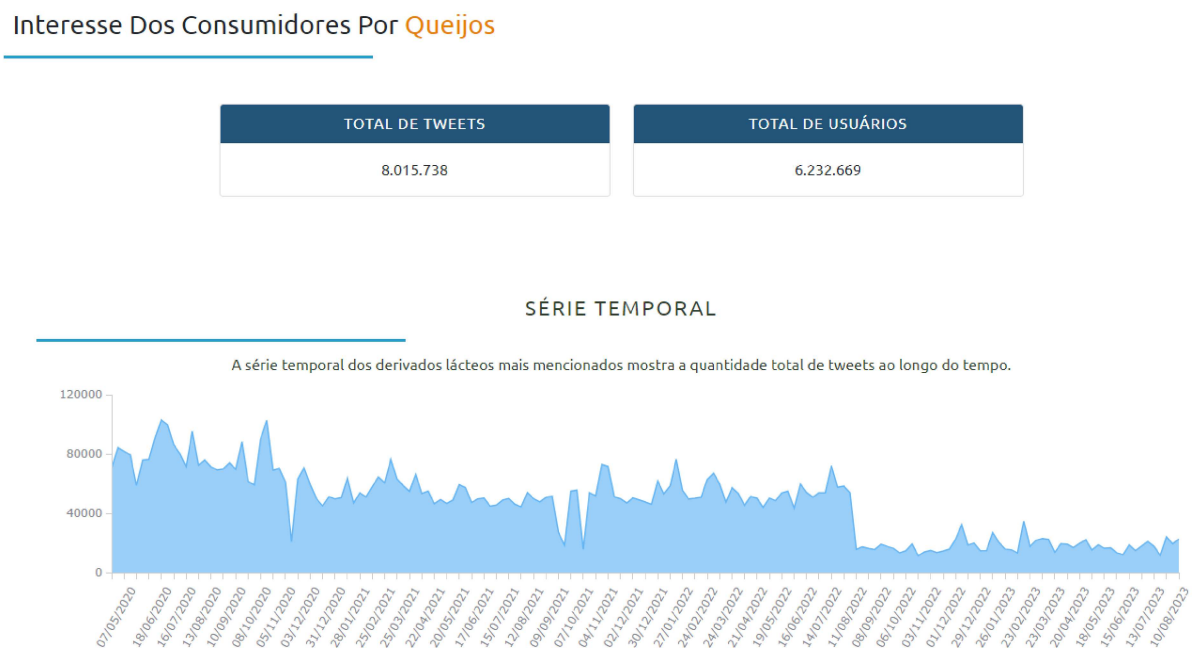


Figura 33 – Gráfico da série temporal da quantidade de postagens do X sobre queijos coletados e processados pelo OC. **Fonte:** Observatório do Consumidor.

A série temporal, como já destacado, desempenha um papel crucial ao fornecer informações valiosas sobre as tendências relacionadas a esse mercado. A análise desta série permite identificar padrões de comportamento ao longo do tempo, tais como picos de interesse ou períodos de menor engajamento. No entanto, é importante destacar que, a partir de 25 de agosto de 2022, houve uma mudança significativa na forma como os dados do X passaram a ser fornecidos pela API oficial. Essa alteração impactou diretamente a coleta de dados, resultando em uma queda observada no número de publicações a partir desse período. Tal redução, contudo, pode não refletir necessariamente uma diminuição real do interesse pelo produto, mas sim uma limitação técnica imposta pela nova estrutura de disponibilização dos dados. Dessa forma, é essencial considerar esses fatores ao interpretar as tendências e comportamentos revelados pela série temporal, a fim de evitar conclusões equivocadas sobre a dinâmica do mercado de queijos.

Dando continuidade às análises, a Figura 34 apresenta duas visualizações cruciais: a primeira ilustra a localização geográfica dos usuários que comentaram sobre queijos, permitindo identificar as regiões de maior interesse e engajamento. A segunda visualização foca na análise de sentimentos aplicada aos comentários coletados, proporcionando uma compreensão mais profunda das percepções dos consumidores.

Observa-se que a região Sudeste foi a que registrou o maior número de publicações sobre queijos, com destaque para os estados de São Paulo, Rio de Janeiro e Minas Gerais, que concentraram o maior número de usuários mencionando o produto em suas postagens. A análise de sentimentos revela que, de modo geral, 75,5% dos comentários são positivos, indicando uma ampla aceitação e experiências satisfatórias dos consumidores em relação

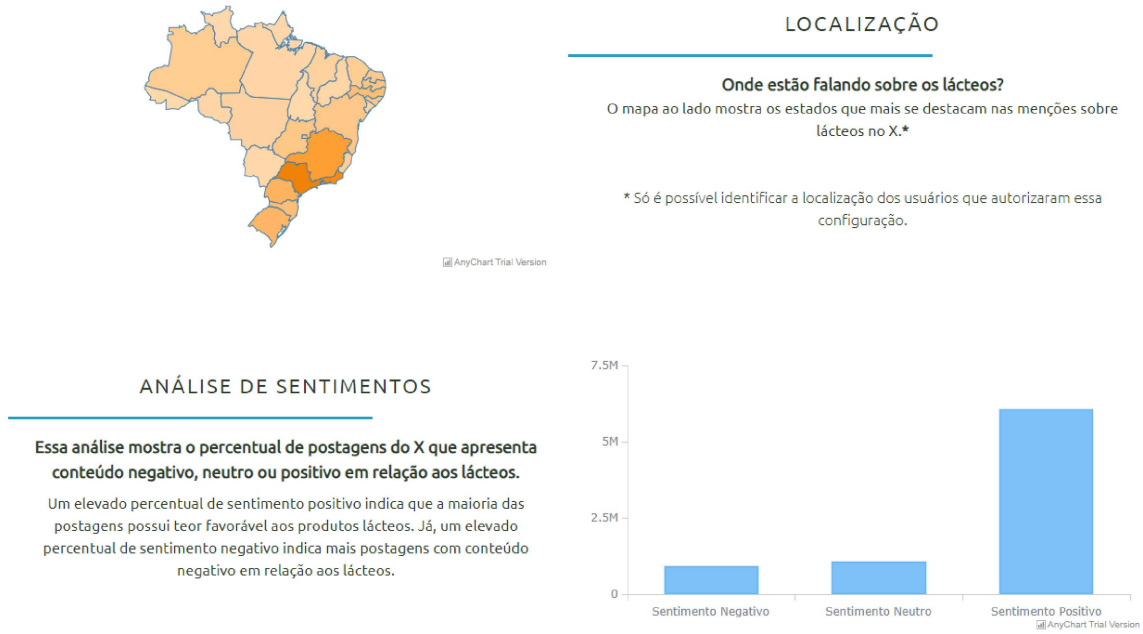


Figura 34 – Gráfico com o mapa do Brasil das regiões que mais publicaram no X sobre queijos e a análise de sentimentos. **Fonte:** Observatório do Consumidor.

aos queijos. Esses resultados sugerem que, ao mencionar ou consumir o produto, os usuários tendem a expressar opiniões favoráveis, evidenciando um elevado nível de satisfação com o produto.

A Figura 35 apresenta duas visualizações adicionais de grande relevância para a análise dos usuários da rede social X que publicaram sobre queijos: o sexo e a geração a que pertencem. Compreender essas características demográficas é fundamental para estimar com maior precisão o perfil do consumidor. Esse entendimento permite explorar preferências específicas desses grupos em relação aos produtos, como os aspectos que mais valorizam e desejam. Essas informações oferecem percepções estratégicas que podem auxiliar o mercado a adaptar suas ofertas, melhorar o atendimento às expectativas desses consumidores e expandir as oportunidades de alcance para novos potenciais clientes.

A análise revela que o sexo masculino foi responsável pela maior parte das publicações sobre queijos, representando 74,68% do total, enquanto o sexo feminino contribuiu com 25,32%. Esse predomínio masculino pode indicar uma maior interação desse grupo com o produto nas redes sociais ou um perfil de consumo mais ativo entre os homens. No que diz respeito à geração dos usuários, a Geração Y, composta por indivíduos nascidos entre 1981 e 1995, destacou-se como a mais engajada. Essa geração, também conhecida como *millennials*, tende a ser um público estratégico, por sua familiaridade com o ambiente digital e seu potencial de influência nas tendências de consumo, o que reforça a importância de direcionar campanhas e estratégias de *marketing* para esse grupo específico.

Analisar as características que um produto possui é de extrema importância para

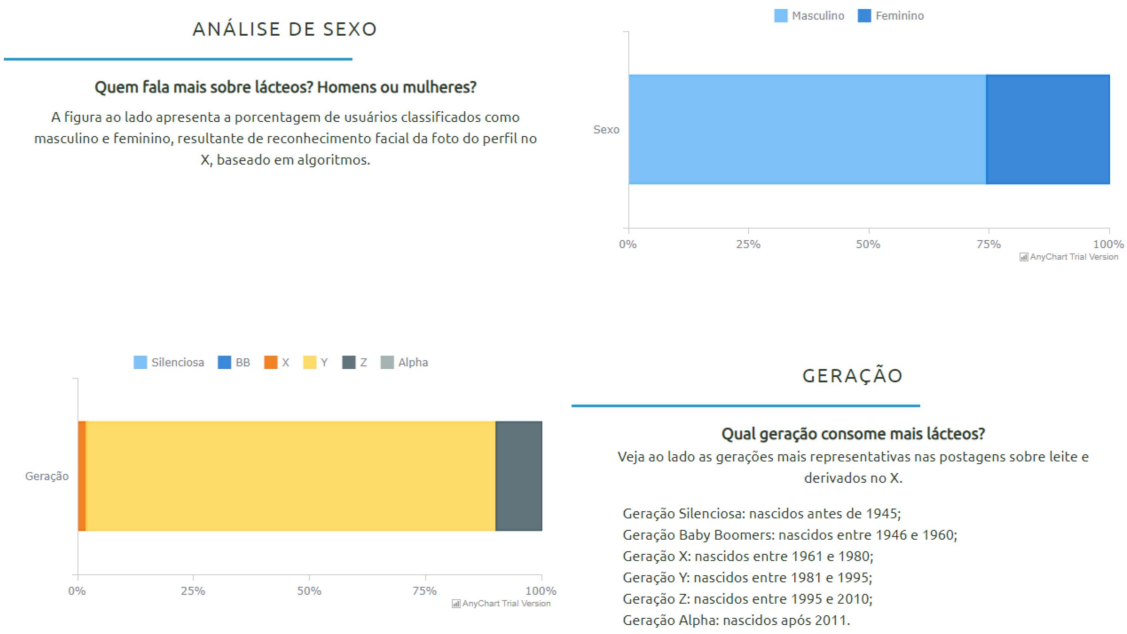


Figura 35 – Gráfico com a análise de sexo e gerações dos usuários do X que mais costumam publicar sobre queijos. **Fonte:** Observatório do Consumidor.

entender como ele é percebido e valorizado pelos consumidores. Para essa avaliação, o OC fornece uma análise detalhada que permite identificar aspectos como a embalagem, informações do produto, composição, métodos de produção e eventuais riscos de contaminação. A Figura 36 ilustra essas informações específicas para os queijos.

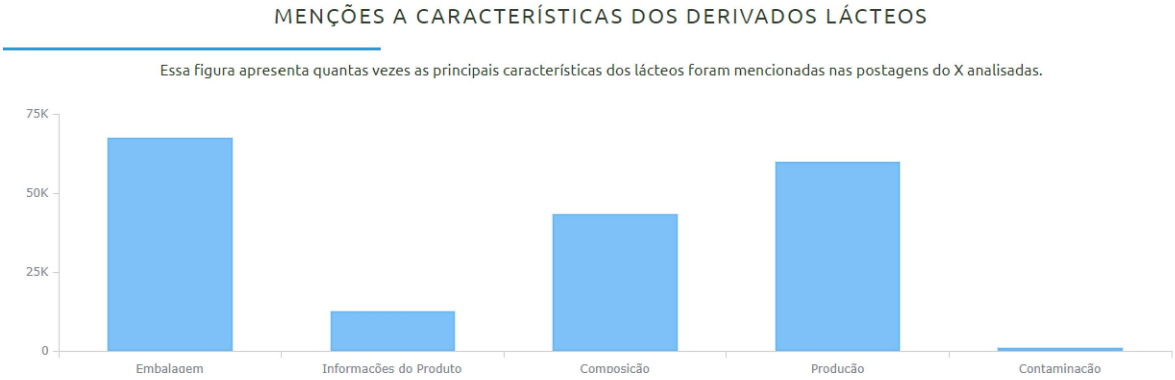


Figura 36 – Gráfico com as principais características dos queijos que os usuários do X mais costumam publicar. **Fonte:** Observatório do Consumidor.

Entre as categorias analisadas, destacam-se aquelas relacionadas à embalagem, produção e composição, que são recorrentes nas publicações sobre queijos. Essas categorias abrangem termos que tratam do *design* e da funcionalidade das embalagens, evidenciando a preferência dos consumidores por praticidade e eficiência na conservação dos produtos. Além disso, são mencionados aspectos como a qualidade e a origem dos ingredientes, os métodos de produção e atributos essenciais, como frescor e segurança alimentar. Essa análise proporciona uma visão abrangente das percepções do público, permitindo identificar

áreas prioritárias para ajustes e melhorias que possam fortalecer a competitividade do produto no mercado.

O OC também investiga os adjetivos mais mencionados nas publicações por meio de uma nuvem de palavras. Essa visualização destaca a frequência com que determinados termos aparecem, além de utilizar cores distintas para representar a polaridade das palavras. Adjetivos com conotação negativa são exibidos em vermelho, neutros em amarelo e positivos em verde. A Figura 37 ilustra essa distribuição em detalhes.

Figura 37 – *Word cloud* com os adjetivos mais frequentes nas publicações dos usuários do X sobre queijos. **Fonte:** Observatório do Consumidor.

É possível observar que palavras como “coalho”, “branco” e “quente” destacam-se por sua alta recorrência nos textos analisados, sugerindo temas centrais nas discussões sobre produtos lácteos. Além disso, identifica-se uma diversidade de percepções, com termos positivos como “bom”, “delicioso”, “maravilhoso” e “perfeito” indicando satisfação e apreço pelos produtos, enquanto expressões negativas como “horrível”, “podre” e “fedorento” refletem críticas e possíveis insatisfações dos consumidores. Essa análise revela tanto os aspectos valorizados quanto os problemáticos na percepção dos consumidores em relação aos produtos lácteos, fornecendo informações valiosas para o direcionamento de estratégias de *marketing* e para a implementação de melhorias na qualidade dos produtos, visando atender às expectativas e necessidades do público. Já na Figura 38, o OC exibe as principais categorias dos alimentos mencionados em conjunto com os queijos nas publicações.

O consumo de queijos é altamente diversificado, como evidenciado pelas frequentes menções nas categorias de “Laticínios”, “Panificados” e “Outros Alimentos Preparados”. A categoria de “Laticínios”, em particular, destacou-se pelo elevado número de menções, o que pode ser explicado pela ampla variedade de queijos disponíveis, além da necessidade

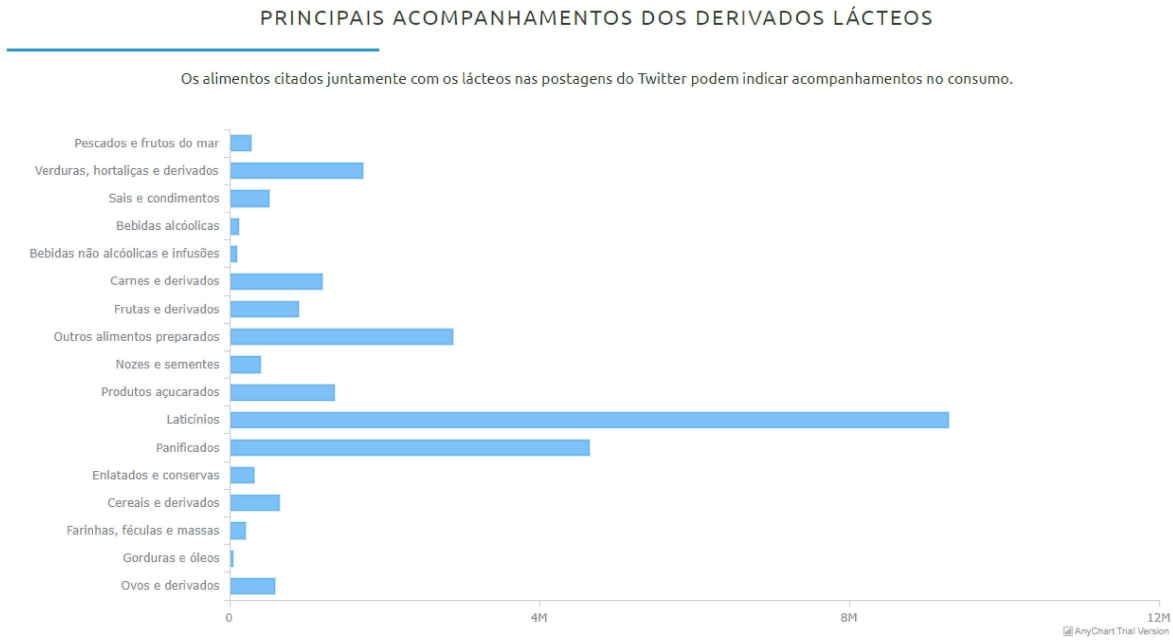


Figura 38 – Categorias dos alimentos mais mencionados em conjunto com os queijos. **Fonte:** Observatório do Consumidor.

de incluir a palavra "queijo" nesse contexto. Ao analisar outras subcategorias de produtos lácteos, como doce de leite, iogurte ou leite condensado, é comum que o queijo esteja presente, seja no consumo associado ou nas menções. Isso justifica a inclusão constante dessa palavra-chave, essencial para uma compreensão mais precisa das discussões sobre o produto.

Por sua vez, a categoria de "Panificados" reflete um hábito comum: o consumo de queijos em combinação com pães, torradas e outros produtos similares. Essa combinação é uma prática amplamente difundida entre os consumidores, o que explica a alta correlação entre as menções a queijos e produtos panificados. Essa análise, portanto, ressalta a versatilidade do queijo na alimentação, sendo consumido tanto como produto principal quanto como acompanhamento em uma ampla variedade de contextos alimentares.

Para aprimorar essa análise, o OC utiliza a nuvem de palavras como um recurso adicional. Nela, são apresentados os produtos alimentares mais mencionados nas postagens sobre queijos, destacados por tamanho e cor. Quanto maior a palavra, maior o número de menções. Cada cor na nuvem possui um significado específico: azul-claro representa alimentos processados, azul-escuro indica alimentos ultraprocessados, laranja refere-se a alimentos *in natura* ou minimamente processados, e amarelo corresponde a ingredientes culinários. A Figura 39 ilustra os resultados relacionados à categoria dos queijos.

A *word cloud* apresentada revela uma clara predominância de alimentos processados e ultraprocessados nas publicações analisadas, com especial destaque para os produtos lácteos. A palavra "queijo" se destaca consideravelmente, o que era esperado, considerando

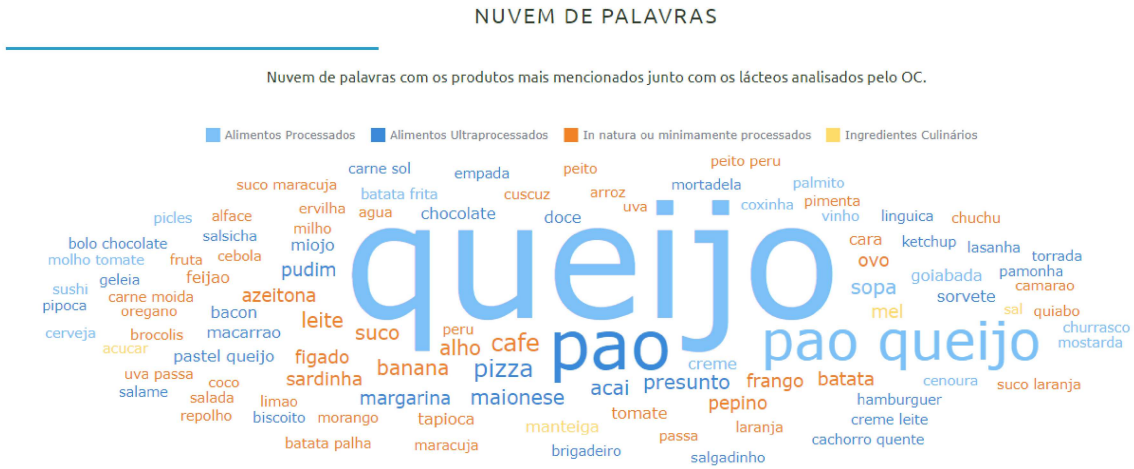


Figura 39 – *Word cloud* com as palavras mais frequentes nas publicações dos usuários do X sobre queijos. **Fonte:** Observatório do Consumidor.

que os dados em análise pertencem a essa categoria. Além disso, termos como "salsicha", "mortadela", "macarrão", "biscoito", "salgadinho" e "*fast food*" (representado por "*hambúrguer*") indicam um consumo elevado de alimentos industrializados, ricos em aditivos e sódio, que se enquadram nas categorias de alimentos processados e ultraprocessados.

Embora os termos "arroz", "feijão", "batata" e "maçã" também estejam presentes, sua frequência é inferior quando comparada aos alimentos processados. A presença de "carne", "frango" e "peixe" é notada, mas a prevalência de cortes processados, como "salsicha" e "mortadela", sugere um maior consumo de carnes industrializadas. Além disso, a menção a "chocolate", "doce", "brigadeiro" e "sobremesa" indica um interesse significativo de alimentos com alto teor de açúcar. Dessa forma, a análise da *word cloud* contribui para identificar padrões e tendências nos hábitos alimentares.

Outra análise que complementa a anterior é a análise comportamental, que visa identificar as principais categorias de ações realizadas pelos usuários da rede social X em relação ao produto queijo. Essa análise busca mapear o comportamento dos consumidores, revelando quais interações e atividades são mais comuns. Na Figura 40, são apresentados os resultados específicos para o queijo.

Entre as categorias analisadas, os verbos relacionados à alimentação e à atividade física foram os que apresentaram o maior número de menções identificadas. No contexto da alimentação, destacaram-se verbos como "preparar", "beber", "comer", "comprar", "desejar" e "fazer", que tiveram a maior influência nas interações dos usuários. Esses verbos refletem tanto o interesse por receitas culinárias quanto o desejo de consumir ou adquirir produtos alimentícios, demonstrando uma relação direta com o comportamento dos consumidores no que diz respeito à alimentação.

A predominância dessas ações confirma as expectativas, dado o foco no universo alimentar, e evidencia um forte engajamento dos usuários em atividades ligadas à expe-

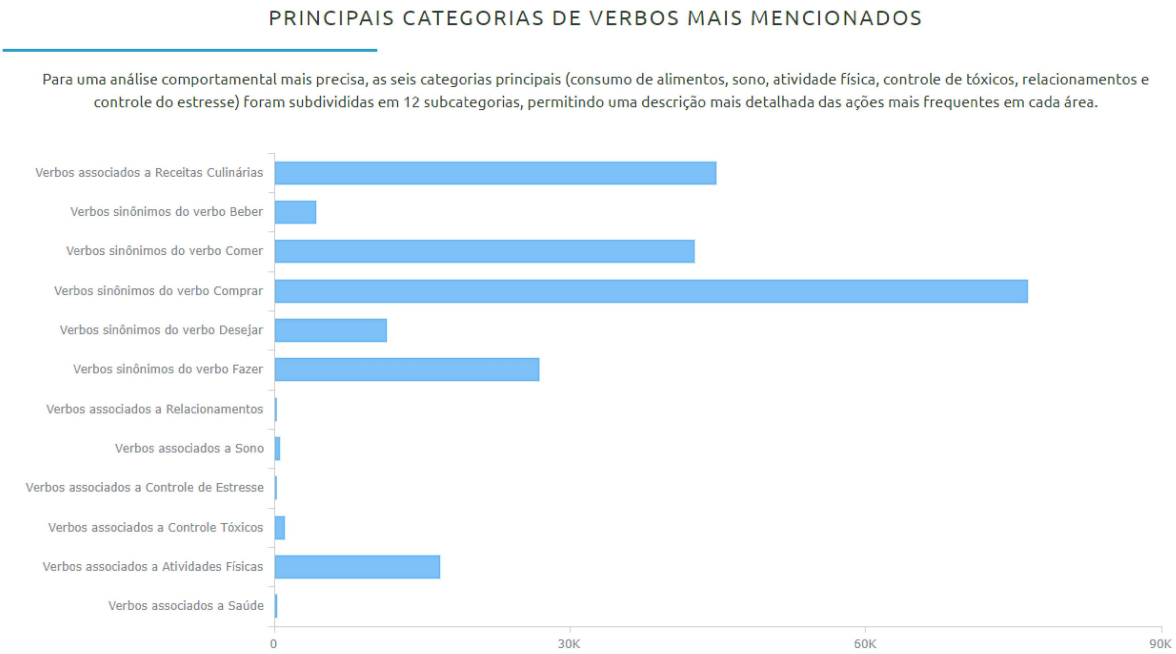


Figura 40 – Categorias de verbos mais mencionados em conjunto com os queijos. **Fonte:** Observatório do Consumidor.

riência gastronômica. Esse engajamento se manifesta de diversas formas, como a busca por inspiração culinária, o compartilhamento de hábitos alimentares, e o planejamento de compras. Além de indicar o interesse dos consumidores por explorar novas receitas e produtos, essas interações também refletem uma conexão emocional com o ato de comer, reforçando a importância dos alimentos no cotidiano dos usuários.

A construção visual do sistema foi cuidadosamente planejada para priorizar a legibilidade e a facilidade de interpretação dos dados. Isso permite que os usuários assimilem rapidamente os principais resultados da pesquisa, contribuindo para uma experiência mais intuitiva e informativa. Outro ponto importante é que as páginas possuem responsividade, ou seja, em qualquer dispositivo o sistema se adapta à tela, o que facilita a navegação dos usuários, proporcionando um ambiente *clean*, fácil e de uso simples para o usuário final.

6 CONTRIBUIÇÕES PARA A LITERATURA

A seguir, estão listados os trabalhos que utilizaram o Observatório do Consumidor, evidenciando sua aplicação e relevância no campo de estudo:

6.1 Artigos Internacionais

- Artigo publicado na *21st International Conference on Intelligent Systems Design and Applications* (ISDA 2021):
  - *Analysis of the Brazilian Artisanal Cheese Market from the Perspective of Social Networks* (NOGUEIRA et al., 2021).
- Artigo publicado na revista *Social Network Analysis and Mining*:
  - *Construction of a Training Dataset for a Sentiment Analysis Model of Dairy Products Tweets in Brazil* (NOGUEIRA et al., 2024).

6.2 Artigos Nacionais

6.2.1 *Workshop* de iniciação científica da Embrapa Gado de Leite

- **Edição de 2021**:
  - Análise exploratória de leite e derivados mais comentados no Twitter e Google Trends (NOGUEIRA et al., 2021b).
  - Quais os tipos de queijos mais comentados no Twitter (SOARES et al., 2021).
  - Análise dos sentimentos expressos na rede social Twitter em relação ao queijo (CAMPOS et al., 2021).
- **Edição de 2022**:
  - Quais os derivados lácteos mais consumidos no Brasil durante a pandemia da COVID-19? (NOGUEIRA et al., 2022b).
  - Análise exploratória do interesse por lácteos orgânicos (SILVA et al., 2022).
  - Impacto da inflação nos *tweets* sobre leite e derivados (FRANCO et al., 2022).
- **Edição de 2024**:
  - Padrões de consumo de queijos no Brasil: Uma abordagem por meio do Observatório do Consumidor. (RODRIGUES et al., 2024).
  - Análise exploratória do interesse por marcas e produtos lácteos no Brasil. (SILVA et al., 2024).

6.2.2 Outros eventos e publicações

- **Workshop Brazilian E-science (BRESCI)**:
  - Uma Arquitetura para a Recomendação de Consumidores de Queijo Artesanal Brasileiro (SOARES et al., 2020).
- **Congresso Brasileiro de Agroinformática 2021** (Agraciado com o Segundo Lugar no Concurso de Teses e Dissertações):
  - Mineração de dados em rede social para avaliação de tendências de consumo do queijo artesanal no Brasil (NOGUEIRA, 2021).
- **SBSI 21: Proceedings of the XVII Brazilian Symposium on Information Systems**:
  - REDIC: *Recommendation of Digital Influencers of Brazilian artisanal cheese* (SOARES et al., 2021).
- **Revista Indústria de Laticínios**:
  - Análise exploratória da imagem dos lácteos em tempos de coronavírus (SIQUEIRA et al., 2020a).
  - O impacto da pandemia no consumo de lácteos no Brasil: uma abordagem das redes sociais (SIQUEIRA et al., 2021b).
- **Artigos no *Milkpoint***:
  - O que mudou no consumo de lácteos com a pandemia? Uma visão das redes sociais (SIQUEIRA et al., 2021b).
  - Tendências de consumo de queijo coalho no Nordeste (NOGUEIRA et al., 2021a).
  - A geografia brasileira do consumo domiciliar de leite (SIQUEIRA et al., 2021a).
  - Os queijos preferidos em Minas Gerais (SIQUEIRA et al., 2024).
- **Anuário do Leite**:
  - Covid-19 e o impacto no consumo de leite e derivados (SIQUEIRA et al., 2021a).
  - Pandemia, leite, queijo e *tweets* (SIQUEIRA et al., 2021c).
- **Revista *The Journal of Engineering and Exact Sciences*** (Agraciado com Menção Honrosa no Simpósio Integrado de Inovação em Tecnologia de Alimentos - UFV, 2022):

  - Mineração de dados em *tweets* para análise do consumo de lácteos no Brasil. (NOGUEIRA et al., 2022a).

- **Livro com textos de artigos publicados**:

  - Na era do consumidor: uma visão do mercado lácteo brasileiro (SIQUEIRA, 2021).

6.3 Registro de *Software*

- **Observatório do Consumidor** (SIQUEIRA et al., 2022): Patente de Programa de Computador, número de registro BR512022002576-0. Este trabalho descreve a criação e a proteção legal do sistema computacional, que tem como objetivo a análise e a interpretação de dados de redes sociais para apoiar pesquisas de mercado.

6.4 *Newsletters*

- Foram publicados 29 temas distintos, acessíveis em https://observatoriodoconsumidor.cnpgl.embrapa.br/newsletter.

6.5 Trabalhos Paralelos (Fora do Contexto de Lácteos)

- **Encontro Nacional de Modelagem Computacional**:

  - *Molecular Dynamics Modeling and Simulation of Tobermorite Under Pre-Salt Conditions* (NOGUEIRA et al., 2022).

- **Revista *Journal of Health Informatics***:

  - Aplicação do *Random Survival Forest* na análise da sobrevida para câncer da mama (CARVALHO et al., 2023).

- **Revista HU Revista da Universidade Federal de Juiz de Fora**:

  - *Social network X and mental health in Brazil in times of coronavirus* (NOGUEIRA et al., 2024b).

7 CONCLUSÃO

Diante do trabalho realizado, conclui-se que os objetivos geral e específicos foram plenamente alcançados. Com base nos princípios de *Business Intelligence*, foi possível desenvolver e implementar o Observatório do Consumidor, um sistema computacional que, por meio de técnicas avançadas de processamento de linguagem natural e aprendizado de máquina, viabiliza a análise de dados de redes sociais para apoiar pesquisas de mercado no setor lácteo brasileiro. Essa ferramenta representa um avanço significativo na compreensão do comportamento do consumidor, oferecendo *insights* estratégicos e embasados em dados.

Foi possível projetar um processo automatizado de coleta, armazenamento e tratamento de dados oriundos das redes sociais com foco em derivados lácteos no Brasil originalmente proposto por NOGUEIRA (2021). A nova versão do sistema passou a incorporar, além da rede social X, dados provenientes do YouTube e do Google Trends, ampliando significativamente o escopo das análises. Com isso, foram incluídas nove novas categorias de produtos, totalizando 10 categorias de derivados lácteos e 293 palavras-chave monitoradas, o que tornou o sistema mais robusto, abrangente e representativo das dinâmicas de consumo do setor lácteo.

Foi modelado um banco de dados relacional escalável, projetado para organizar grandes volumes de dados sobre produtos lácteos de forma eficiente. Com a expansão do escopo da ferramenta para incluir novos produtos na coleta, o banco passou por ajustes estruturais que aumentaram sua capacidade de armazenamento e flexibilidade. Para garantir um tempo de resposta ágil aos usuários finais, foram implementadas tabelas de resultados otimizadas, permitindo consultas mais rápidas e melhorando significativamente o desempenho da interface do sistema computacional.

Para aprimorar a análise de sentimentos diante do aumento no volume de dados e da diversificação das categorias de produtos, foi construído e validado um *corpus* de treinamento em português do Brasil específico para o setor de derivados lácteos. O conjunto de dados, balanceado e abrangente, contempla todas as categorias analisadas, permitindo maior representatividade nas avaliações. Com base no modelo inicial proposto por NOGUEIRA (2021), o trabalho de NOGUEIRA et al. (2024) promoveu uma expansão significativa da arquitetura, adotando uma abordagem mais robusta que resultou em classificações mais precisas dos sentimentos expressos pelos consumidores sobre os produtos lácteos.

Para além da análise de sentimentos, foram aplicadas técnicas de mineração textual e linguística computacional com o objetivo de identificar padrões linguísticos relacionados a características, hábitos e preferências de consumo no setor de derivados lácteos. A partir do desenvolvimento de algoritmos especializados, tornou-se possível extrair relações sintáticas e semânticas presentes nos textos das publicações, permitindo uma compreensão mais

profunda do comportamento dos consumidores. Esse avanço metodológico representa uma evolução importante em relação à abordagem inicial de NOGUEIRA (2021), ampliando a capacidade analítica do sistema e sua aplicabilidade no contexto brasileiro.

A avaliação da eficácia do uso de reconhecimento facial e da análise de imagens de perfil público foi incorporada como apoio à identificação de perfis de consumidores, sempre em conformidade com os critérios éticos e legais da pesquisa. A partir da aplicação de modelos previamente treinados, foi possível extrair características como sexo e faixa etária dos usuários que publicaram sobre produtos lácteos nas redes sociais. Essa abordagem complementou a análise textual ao enriquecer o perfil demográfico dos consumidores, oferecendo uma visão mais ampla e estratégica para o entendimento do mercado.

Foi desenvolvido um novo sistema *web* com processamento de consultas em banco de dados otimizado e visualização interativa, consolidando e organizando as informações obtidas pelo Observatório do Consumidor. A interface, projetada de forma intuitiva, permite a apresentação clara e estratégica dos dados coletados e analisados, facilitando a tomada de decisão por parte de pesquisadores, indústria e formuladores de políticas públicas no setor lácteo brasileiro. O sistema está disponível em: https://observatoriodoconsumidor.cnpgl.embrapa.br.

O sistema computacional desenvolvido foi validado por meio de estudos de caso reais no setor de lácteos no Brasil, comprovando sua eficácia em contextos práticos. As funcionalidades implementadas foram projetadas para atender às necessidades de pesquisadores e *stakeholders*, oferecendo uma alternativa eficiente aos desafios enfrentados pelas abordagens tradicionais de pesquisa de mercado, como tempo de resposta, volume e diversidade dos dados, representatividade e custos operacionais. Com isso, o Observatório do Consumidor se estabelece como uma ferramenta estratégica, fornecendo acesso contínuo e detalhado a informações sobre hábitos de consumo, tendências de mercado e comportamento de usuários nas redes sociais em relação aos derivados lácteos.

Além de ampliar o conhecimento sobre o comportamento do consumidor no setor, o Observatório do Consumidor fortalece a tomada de decisão no mercado, permitindo comparações mais precisas entre os produtos analisados e oferecendo um ambiente ágil para a obtenção de *insights* estratégicos. Essa integração entre tecnologia e análise de dados representa um avanço significativo, tornando o Observatório do Consumidor um recurso essencial para impulsionar a inovação e a competitividade no setor lácteo.

7.1 Limitações da Ferramenta

Apesar das diversas funcionalidades já implementadas, a ferramenta desenvolvida ainda apresenta algumas limitações relevantes. Um dos principais desafios está na coleta de dados em redes sociais, especialmente após as mudanças na política de acesso da plataforma X. Desde 24/08/2023, a adoção de diretrizes mais restritivas pela nova gestão

dificultou significativamente a obtenção de informações, inviabilizando o uso contínuo e gratuito da API.

Atualmente, embora tenha havido ajustes na política de compartilhamento, as restrições permanecem expressivas. A versão gratuita da API permite apenas 100 requisições por mês - um volume irrisório para análises consistentes. Já o plano mais acessível, denominado *Basic*, custa aproximadamente US$ 200 por mês e oferece até 10.000 requisições, enquanto o plano Pro, com um custo em torno de US$ 5.000 mensais, disponibiliza um volume substancialmente maior.

Diante desse cenário, a limitação no acesso a um volume adequado de dados compromete a realização de análises robustas. Como alternativa, é fundamental ampliar a coleta e a análise de informações em outras redes sociais, assegurando que o Observatório do Consumidor continue cumprindo seu papel estratégico na produção de conhecimento sobre o mercado e o comportamento dos consumidores.

7.2 Trabalhos Futuros

O Observatório do Consumidor apresenta um potencial expressivo para futuras expansões, tanto em escopo quanto em profundidade analítica. Sua evolução pode ocorrer por meio da incorporação de novas abordagens metodológicas e da ampliação das fontes de dados, fortalecendo sua capacidade de gerar *insights* estratégicos e relevantes para o setor lácteo.

Uma linha de pesquisa promissora envolve a identificação e categorização das emoções expressas pelos consumidores após o consumo de produtos lácteos e ao compartilharem suas opiniões nas redes sociais sobre esses produtos. Essa abordagem, além da análise da polaridade dos sentimentos já realizada pelo OC, permitiria uma compreensão mais aprofundada das reações do público em relação a marcas, categorias específicas e campanhas de *marketing*. A adoção dessa metodologia não só enriqueceria as análises realizadas pelo Observatório do Consumidor, mas também consolidaria ainda mais a ferramenta como um referencial inovador no estudo do mercado lácteo e do comportamento do consumidor nas redes sociais.

Outra vertente de desenvolvimento está na análise de dados multimodais. A integração de informações textuais com imagens, vídeos e áudios extraídos de redes sociais pode enriquecer a compreensão das dinâmicas de consumo. A combinação dessas diversas fontes de informação permitirá identificar padrões de comportamento com maior precisão e capturar nuances que análises exclusivamente textuais poderiam deixar escapar. Dessa forma, o uso de abordagens multimodais desponta como uma oportunidade relevante para enriquecer a análise de mercado e gerar *insights* mais completos e estratégicos.

Por fim, o Observatório do Consumidor também pode ser adaptado para o mo-

nitoramento de outros setores do agronegócio, como carnes, grãos ou outros tipos de produtos, replicando a arquitetura já desenvolvida com ajustes específicos para novas cadeias produtivas. Tal ampliação consolidaria o Observatório como uma plataforma versátil de inteligência de mercado baseada em dados sociais, com aplicação direta em políticas públicas, inovação tecnológica e estratégias comerciais em múltiplos segmentos do agronegócio brasileiro.

REFERÊNCIAS

AL-DYANI, Wafa Zubair; AHMAD, Farzana Kabir; KAMARUDDIN, Siti Sakira. Adaptive binary bat and markov clustering algorithms for optimal text feature selection in news events detection model. **IEEE Access**, v. 10, p. 85655–85676, 2022. ISSN 2169-3536.

AL-YASIRI, Essam; AL-AZAWEI, Ahmed. Improving arabic sentiment analysis on social media: A comparative study on applying different pre-processing techniques. v. 8, p. 3150–3157, 06 2019.

ALHARBI, Fatmah; KHAN, Muhammad Badruddin. Identifying comparative opinions in arabic text in social media using machine learning techniques. **SN Applied Sciences**, v. 1, p. 1–13, 2019.

ALIJANI, Shadi; TANHA, Jafar; MOHAMMADKHANLI, Leyli. An ensemble of deep learning algorithms for popularity prediction of flickr images. **Multimedia Tools and Applications**, v. 81, 01 2022.

AMORIM, Marta; BORTOLOTI, Frederico D.; CIARELLI, Patrick M.; SALLES, Evandro O. T.; CAVALIERI, Daniel C. Novelty detection in social media by fusing text and image into a single structure. **IEEE Access**, v. 7, p. 132786–132802, 2019. ISSN 2169-3536.

AZATISOFTWARE. Automated data labeling with machine learning. 2019. Accessed: 2025-01-02. Disponível em: https://azati.ai/automated-data-labeling-with-machine-learning.

BAKKEN, S. S.; SURASKI, Z.; SCHMID, E. **PHP Manual: Volume 1**. [S.l.]: iUniverse, Incorporated, 2000.

BAPPON, Suborno Deb; IQBAL, Asif. Classification of tourism reviews from bengali texts using multinomial naïve bayes. In: **2022 25th International Conference on Computer and Information Technology (ICCIT)**. [S.l.: s.n.], 2022. p. 270–275.

BARABBA, T.; ZALTAMAN, P. Hearing the voice of the market. **Harvard Business School Press**, 1991.

BARROS, A. P. **A Importância da Pesquisa de Mercado e Inovação para Empresas de Sonorização e Iluminação de Eventos do Distrito Federal: Um Estudo de Caso da Marc Systems**. 2010. Disponível no Repositório do UniCEUB. Disponível em: https://repositorio.uniceub.br/jspui/bitstream/235/9199/1/20800101.pdf.

BATISTA, Gustavo; BAZZAN, Ana; MONARD, Maria-Carolina. Balancing training data for automated annotation of keywords: a case study. In: . [S.l.: s.n.], 2003. p. 10–18.

BATISTA, Gustavo E. A. P. A.; PRATI, Ronaldo C.; MONARD, Maria Carolina. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explor. Newsl.**, Association for Computing Machinery, New York, NY, USA, v. 6, n. 1, p. 20–29, jun 2004. ISSN 1931-0145. Disponível em: https://doi.org/10.1145/1007730.1007735.

BENIS, Arriel; CHATSUBI, Anat; LEVNER, Eugene; ASHKENAZI, Shai. Change in threads on twitter regarding influenza, vaccines, and vaccination during the covid-19 pandemic: Artificial intelligence–based infodemiology study. **JMIR Infodemiology**, v. 1, n. 1, p. e31983, Oct 2021. ISSN 2564-1891. Disponível em: https://infodemiology.jmir.org/2021/1/e31983.

BHATNAGAR, Sarvesh; CHOUBEY, Nitin. Making sense of tweets using sentiment analysis on closely related topics. 04 2021.

BRANDÃO, Julianda Crislaine da Silva. **SISTEMAS DE INFORMAÇÃO INTEGRADOS: UMA ANÁLISE SOBRE O PROCESSO DE DESENVOLVIMENTO E IMPLANTAÇÃO DE BUSINESS INTELLIGENCE EM UMA DISTRIBUIDORA DE MEDICAMENTOS**. 2016. Dissertação (B.S. thesis), 2016.

BRITO, Edeleon Marcelo Nunes de. Mineração de textos: Detecção automática de sentimentos em comentários nas mídias sociais. **Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento**, v. 6, 2017. Disponível em: https://api.semanticscholar.org/CorpusID:157610558.

CAMACHO, P. A. F. **Sistema de recomendação em real-time para reserva de transfers**. 2020. Dissertação (Dissertação de mestrado) — Iscte - Instituto Universitário de Lisboa, 2020. Repositório do Iscte. Disponível em: http://hdl.handle.net/10071/22131.

CAMBRIA, Erik; SCHULLER, Björn; XIA, Yunqing; HAVASI, Catherine. New avenues in opinion mining and sentiment analysis. **Intelligent Systems, IEEE**, v. 28, p. 15–21, 03 2013.

CAMPOS, E.; SOARES, N. D.; NOGUEIRA, T.; SIQUEIRA, K. B.; MORAES, E. Análise dos sentimentos expressos na rede social twitter em relação ao queijo. In: **Anais do Workshop de iniciação científica da Embrapa Gado de Leite**. Juiz de Fora, MG, Brasil: Embrapa Gado de Leite, 2021.

CARVALHO, Daniela Schimitz de; NOGUEIRA, Thallys da Silva; GOLIATT, Priscila Vanessa Zabala Caprile. Aplicação do random survival forest na análise da sobrevida para câncer da mama. **Journal of Health Informatics**, v. 15, n. Especial, jul. 2023. Disponível em: https://www.jhi.sbis.org.br/index.php/jhi-sbis/article/view/1113.

CAVALCANTE, Paulo Emílio Costa; BARBOSA, Yuri de Almeida Malheiros. Um dataset para análise de sentimentos na língua portuguesa. **Departamento de Ciências Exatas - Universidade Federal da Paraíba (UFPB)**, 2017. Rio Tinto - PB - Brasil.

CHANG, Yung-Chun; KU, Chih-Hao; NGUYEN, Duy-Duc Le. Predicting aspect-based sentiment using deep learning and information visualization: The impact of covid-19 on the airline industry. **Information Management**, v. 59, n. 2, p. 103587, 2022. ISSN 0378-7206. Disponível em: https://www.sciencedirect.com/science/article/pii/S0378720621001610.

CHAWLA, N.; BOWYER, K.; HALL, Lawrence O.; KEGELMEYER, W. Philip. Smote: Synthetic minority over-sampling technique. **ArXiv**, abs/1106.1813, 2002.

CHEN, Keyu; FEI, Cheng; BI, Ziqian; LIU, Junyu; PENG, Benji; ZHANG, Sen; PAN, Xuanhe; XU, Jiawei; WANG, Jinlang; YIN, Caitlyn Heqi; ZHANG, Yichao; FENG, Pohsun; WEN, Yizhu; WANG, Tianyang; LI, Ming; REN, Jintao; NIU, Qian; CHEN, Silin; HSIEH, Weiche; YAN, Lawrence K. Q.; LIANG, Chia Xin; XU, Han; TSENG,

Hong-Ming; SONG, Xinyuan; LIU, Ming. **Deep Learning and Machine Learning – Natural Language Processing: From Theory to Application**. 2024. Disponível em: https://arxiv.org/abs/2411.05026.

CHERNYAEV, Aleksandr; SPRYISKOV, Alexey; IVASHKO, Alexander; BIDULYA, Yuliya. A rumor detection in russian tweets. In: KARPOV, Alexey; POTAPOVA, Rodmonga (Ed.). **Speech and Computer**. Cham: Springer International Publishing, 2020. p. 108–118. ISBN 978-3-030-60276-5.

COVINGTON, M. Nlp for prolog programmers. **Prentice-Hall**, 1994.

DATAREPORTAL. **Digital 2018: Q4 Global Digital Statshot**. 2018. Disponível em: https://datareportal.com/reports/digital-2018-q4-global-digital-statshot.

DATAREPORTAL. **Digital 2022 Global Digital Overview**. 2022. Disponível em: https://datareportal.com/reports/digital-2022-global-overview-report.

DATAREPORTAL. **Social Media Users: The Latest Stats and Trends**. 2024. Accessed: 2025-02-01. Disponível em: https://datareportal.com/social-media-users.

DESMOND, Michael; DUESTERWALD, Evelyn; BRIMIJOIN, Kristina; BRACHMAN, Michelle; PAN, Qian. Semi-automated data labeling. In: ESCALANTE, Hugo Jair; HOFMANN, Katja (Ed.). **Proceedings of the NeurIPS 2020 Competition and Demonstration Track**. PMLR, 2021. (Proceedings of Machine Learning Research, v. 133), p. 156–169. Disponível em: https://proceedings.mlr.press/v133/desmond21a.html.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: https://arxiv.org/abs/1810.04805.

DIGITAL APP. **O que é Widget: definição e exemplos práticos**. 2025. Acessado em: 10 jan. 2025. Disponível em: https://digital.app.br/glossario/o-que-e-widget-definicao-e-exemplos-praticos/.

DUAN, Lei; XU, Lida. Business intelligence for enterprise systems: A survey. **IEEE Transactions on Industrial Informatics**, v. 8, n. 3, p. 679–687, 2012.

D'ANDREA, A.; FERRI, F.; GRIFONI, P.; GUZZO, T. Approaches, tools and applications for sentiment analysis and implementation. **International Journal of Computer Applications**, v. 125, n. 3, p. 26–33, 2015.

EBERHARD DAVID M., Gary F. Simons; FENNIG, Charles D. Ethnologue: Languages of the world. Texas: SIL International, 2022. Disponível em: http://www.ethnologue.com.

EKER, Sibel; GARCIA, David; VALIN, Hugo; RUIJVEN, Bas van. Using social media audience data to analyse the drivers of low-carbon diets. **Environmental Research Letters**, v. 16, 07 2021.

EL-HASNONY, Ibrahim M.; ELZEKI, Omar M.; ALSHEHRI, Ali; SALEM, Hanaa. Multi-label active learning-based machine learning model for heart disease prediction. **Sensors**, v. 22, n. 3, 2022. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/22/3/1184.

FAN, Hong; DU, Wu; DAHOU, Abdelghani; EWEES, Ahmed A.; YOUSRI, Dalia; ELAZIZ, Mohamed Abd; ELSHEIKH, Ammar H.; ABUALIGAH, Laith; AL-QANESS, Mohammed A. A. Social media toxicity classification using deep learning: Real-world application uk brexit. **Electronics**, v. 10, n. 11, 2021. ISSN 2079-9292. Disponível em: https://www.mdpi.com/2079-9292/10/11/1332.

FRANCO, A. L.; SILVA, D. H. C.; NOGUEIRA, T. S.; SIQUEIRA, K. B.; GOLIATT, P. V. Z. C. Impacto da inflação nos tweets sobre leite e derivados. **XXVI Workshop de Iniciação Científica da Embrapa Gado de Leite - 145 PIBIC/CNPq 2021/2022**, 2022. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1148824/1/Impacto-da-inflacao-nos-tweets-sobre-leite-e-derivados.pdf.

FREDRIKSSON, Teodor; MATTOS, David Issa; BOSCH, Jan; OLSSON, Helena. Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In: **International Conference on Product-Focused Software Process Improvement**. [S.l.: s.n.], 2020. p. 202–216.

GITHUB. **GitHub: Where the World Builds Software**. 2024. https://github.com. Acesso em: 8 fev. 2024.

GONZALEZ, Leandro de Azevedo. Regressão logística e suas aplicações. 2018.

GOOGLE. **Google API Client Library for Python**. 2013. https://github.com/googleapis/google-api-python-client. Accessed: 2025-01-07.

GUPTA, Itisha; JOSHI, Nisheeth. Tweet normalization: A knowledge based approach. In: **2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)**. [S.l.: s.n.], 2017. p. 157–162.

HACKERNOON. **Crowdsourcing Data Labeling for Machine Learning Projects**. 2020. Accessed: 2025-01-02. Disponível em: https://hackernoon.com/crowdsourcing-data-labeling-for-machine-learning-projects-a-how-to-guide-cp6h32nd.

HAN, Wenjuan; JIANG, Yong; NG, Hwee Tou; TU, Kewei. **A Survey of Unsupervised Dependency Parsing**. 2020. Disponível em: https://arxiv.org/abs/2010.01535.

HAVRLANT, Lukáš; KREINOVICH, Vladik. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). **International Journal of General Systems**, Taylor & Francis, v. 46, n. 1, p. 27–36, 2017.

HE, Haibo; BAI, Yang; GARCIA, Edwardo A.; LI, Shutao. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: **2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)**. [S.l.: s.n.], 2008. p. 1322–1328. ISSN 2161-4407.

HNAIF, Adnan; KANAN, Emran; KANAN, Tarek. Sentiment analysis for arabic social media news polarity. **Intelligent Automation Soft Computing**, v. 28, p. 107–119, 01 2021.

HONNIBAL, Matthew; MONTANI, Ines. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 2017.

IBGE. **Pesquisa de Orçamentos Familiares 2017-2018: análise do consumo alimentar pessoal no Brasil.** Rio de Janeiro: [s.n.], 2020. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101742.pdf.

IBGE, Instituto Brasileiro de Geografia e Estatística. **Pesquisa de orçamentos familiares 2008-2009 : análise do consumo alimentar pessoal no Brasil. Coordenação de Trabalho e Rendimento.** 2011. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv50063.pdf.

IBGE/SIDRA. **IPCA - Índice de Preços ao Consumidor Amplo**. 2022. https://sidra.ibge.gov.br/pesquisa/snipc/ipca/. Acesso em: jun. 17, 2022.

IBM. **O que é rotulagem de dados?** 2024. https://www.ibm.com/br-pt/topics/data-labeling. Acesso em: 20 dez. 2024.

IFOAM GENERAL ASSEMBLY. **International Federation Of Organic Agriculture Movements. Definition of Organic Agriculture**. 2008. Disponível em: https://www.ifoam.bio/why-organic/organic-landmarks/definition-organic. Acesso em: 15 de Fevereiro de 2025. Disponível em: https://www.ifoam.bio/why-organic/organic-landmarks/definition-organic.

INDURKHYA, N.; DAMERAU, F.J. **Handbook of Natural Language Processing (2nd ed.)**. [S.l.]: Chapman and Hall/CRC, 2010.

JAFARZADEH, Hamed; PAULEEN, David; ABEDIN, Ehsan; WEERASINGHE, Kasuni; TASKIN, Nazim; COşKUN, Mustafa. Making sense of covid-19 over time in new zealand: Assessing the public conversation using twitter. **PLOS ONE**, v. 16, p. e0259882, 12 2021.

JONATHAN, Bern; PUTRA, Panca Hadi; RULDEVIYANI, Yova. Observation imbalanced data text to predict users selling products on female daily with smote, tomek, and smote-tomek. In: **2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)**. [S.l.: s.n.], 2020. p. 81–85.

KANAN, Tarek; KANAAN, Ghassan G.; AL-SHALABI, Riyad; ALDAAJA, Amal. Offensive language detection in social networks for arabic language using clustering techniques. In: . [S.l.: s.n.], 2021.

KANG, Daekook; PARK, Yongtae. Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and vikor approach. **Expert Systems with Applications**, v. 41, n. 4, Part 1, p. 1041–1050, 2014. ISSN 0957-4174. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417413006027.

KHAN, Qasim; KALBUS, Edda; ZAKI, Nazar; MOHAMED, Mohamed M. Utilization of social media in floods assessment using data mining techniques. **PLoS ONE**, v. 17, p. e0267079, 04 2022.

KOULOUMPIS, E.; WILSON, T.; MOORE, J. D. Twitter sentiment analysis: The good the bad and the omg! In: **Proceedings of the Fifth International Conference on Weblogs and Social Media**. Barcelona, Catalonia, Spain: AAAI Press, 2011. p. 538–541. Disponível em: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2857.

KPMG. **Global Organic Milk Production Market Report.** 2018. Disponível em: https://ciorganicos.com.br/wp-content/uploads/2020/09/global-organic-milk-production-market-report.pdf. Acesso em: 23 de julho de 2022.

KULMIZEV, Artur. **The Search for Syntax : Investigating the Syntactic Knowledge of Neural Language Models Through the Lens of Dependency Parsing**. 2023. 101 p. Tese (Doutorado) — Uppsala University, Department of Linguistics and Philology, 2023.

KURNIA, Parama Fadli; SUHARJITO. Business intelligence model to analyze social media information. **Procedia Computer Science**, v. 135, p. 5–14, 2018. ISSN 1877-0509. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050918314303.

LAMPLE, Guillaume; CONNEAU, Alexis; DENOYER, Ludovic; RANZATO, Marc'Aurelio. **Unsupervised Machine Translation Using Monolingual Corpora Only**. 2018. Disponível em: https://arxiv.org/abs/1711.00043.

LEMAîTRE, Guillaume; NOGUEIRA, Fernando; ARIDAS, Christos K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research**, v. 18, n. 17, p. 1–5, 2017. Disponível em: http://jmlr.org/papers/v18/16-365.html.

LINKE, Leonardo Lavalhos. O impacto da implementação de ferramentas de business intelligence na gestão hospitalar: desafios e potencialidades. Universidade Federal de Santa Maria, 2024.

LIU, Bing. Sentiment analysis and opinion mining. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers {LLC}, v. 5, n. 1, p. 1–167, maio 2012. Disponível em: http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016.

LIU, Bing. **Sentiment Analysis: Mining Opinions, Sentiments, and Emotions**. [S.l.: s.n.], 2015. 1-367 p. ISBN 9781107017894.

LIU, Bing; ZHANG, Lei. A survey of opinion mining and sentiment analysis. In: _____. **Mining Text Data**. Boston, MA: Springer US, 2012. p. 415–463. ISBN 978-1-4614-3223-4. Disponível em: https://doi.org/10.1007/978-1-4614-3223-4_13.

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: **In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics**. [S.l.: s.n.], 2002.

LYU, Joanne Chen; LULI, Garving K. Understanding the public discussion about the centers for disease control and prevention during the covid-19 pandemic using twitter data: Text mining analysis study. **J Med Internet Res**, v. 23, n. 2, p. e25108, Feb 2021. ISSN 1438-8871. Disponível em: http://www.jmir.org/2021/2/e25108/.

MAGLAPUZ, Sharmaine Justyne Ramos; LACATAN, Luisito Lolong. Text categorization of filipino tweets using na¨ıve byes algorithm. In: **2021 The 4th International Conference on Machine Learning and Machine Intelligence**. New York, NY, USA:

Association for Computing Machinery, 2022. (MLMI'21), p. 44–49. ISBN 9781450384247. Disponível em: https://doi.org/10.1145/3490725.3490732.

MARCIN, Junczys-Dowmunt; ROMAN, Grundkiewicz; TOMASZ, Dwojak; HIEU, Hoang; KENNETH, Heafield; TOM, Neckermann; FRANK, Seide; ULRICH, Germann; FIKRI, Aji Alham; NIKOLAY, Bogoychev; T., Martins André F.; ALEXANDRA, Birch. Marian: Fast neural machine translation in C++. In: **Proceedings of ACL 2018, System Demonstrations**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 116–121. Disponível em: https://aclanthology.org/P18-4020.

MATHIEU, E.; RITCHIE, H.; ORTIZ-OSPINA, E.; ROSER, M.; HASELL, J.; APPEL, C.; GIATTINO, C.; RODéS-GUIRAO, L. A global database of covid-19 vaccinations. **Nature human behaviour**, p. 947–953, 2021.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). **Proceedings of the 9th Python in Science Conference**. [S.l.: s.n.], 2010. p. 56 – 61.

MEMON, Atia Bano; SOOTAHAR, Dileep Kumar; LUHANA, Kirshan Kumar; MEYER, Kyrill. A corpus-based real-time text classification and tagging approach for social data. **Frontiers in Computer Science**, v. 6, 2024. ISSN 2624-9898. Disponível em: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1294985.

MINISTÉRIO DA SAÚDE. **Guia alimentar para a população brasileira.** 2014. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-brasil/publicacoes-para-promocao-a-saude/guia_alimentar_populacao_brasileira_2ed.pdf/@@download/file/guia_alimentar_populacao_brasileira_2ed.pdf.

MO, Chen; YIN, Jingjing; FUNG, Isaac Chun-Hai; TSE, Zion Tsz Ho. Aggregating twitter text through generalized linear regression models for tweet popularity prediction and automatic topic classification. **European Journal of Investigation in Health, Psychology and Education**, v. 11, n. 4, p. 1537–1554, 2021. ISSN 2254-9625. Disponível em: https://www.mdpi.com/2254-9625/11/4/109.

MUGHAID, Ala; OBEIDAT, Ibrahim; HAWASHIN, Bilal; ALZU'BI, Shadi; AQEL, Darah. A smart geo-location job recommender system based on social media posts. In: **2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)**. [S.l.: s.n.], 2019. p. 505–510.

MW, Kearney. rtweet: Collecting and analyzing twitter data. **Journal of Open Source Software**, v. 4, n. 42, p. 1829, 2019.

MYSQL AB. **MySQL: The World's Most Popular Open Source Database**. [S.l.], 2024. Disponível em: https://www.mysql.com. Acesso em: 8 fev. 2024.

NARAYANAN, Ramanathan; LIU, Bing; CHOUDHARY, Alok. Sentiment analysis of conditional sentences. In: KOEHN, Philipp; MIHALCEA, Rada (Ed.). **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing**. Singapore: Association for Computational Linguistics, 2009. p. 180–189. Disponível em: https://aclanthology.org/D09-1019/.

NEPA, Núcleo de Estudos e Pesquisas em Alimentação. **Tabela Brasileira de Composição de Alimentos – TACO. 4ª edição revisada e ampliada.** 2011. Disponível em: https://www.cfn.org.br/wp-content/uploads/2017/03/taco_4_edicao_ampliada_e_revisada.pdf.

NETO, Luiz Gonzaga da Costa; CAMPOS, Fernando Celso de. Oportunidades de aplicações de business intelligence no contexto da indústria 4.0: revisão sistemática da literatura 2015-2020. **Exacta**, v. 21, n. 2, p. 503–519, jun. 2023. Disponível em: https://periodicos.uninove.br/exacta/article/view/19525.

NG, Ka Chung; TANG, Jie; LEE, Dongwon. The effect of platform intervention policies on fake news dissemination and survival: An empirical examination. **Journal of Management Information Systems**, M.E. Sharpe Inc., v. 38, n. 4, p. 898–930, jan 2022. ISSN 0742-1222. Publisher Copyright: © 2021 The Author(s). Published with license by Taylor Francis Group, LLC.

NOGUEIRA, TS; SIQUEIRA, KB; SOARES, ND; CAMPOS, EW; MORAES, EAP; VILLELA, RMMB; DAVID, JMN; GOLIATT, PVZC; PIRES, M; DINIZ, FH et al. Tendências de consumo de queijo coalho no nordeste. **MilkPoint**, mar 2021.

NOGUEIRA, T.; SOARES, N. D.; CAMPOS, E.; SIQUEIRA, K. B.; GOLIATT, P. C. Análise exploratória de leite e derivados mais comentados no twitter e google trends. In: **Anais do Workshop de iniciação científica da Embrapa Gado de Leite**. Juiz de Fora, MG, Brasil: Embrapa Gado de Leite, 2021.

NOGUEIRA, Thallys da Silva; SARTE, Laisa Marcorela Andreoli; GOLIATT, Priscila Vanessa Zabala Capriles. Social network x and mental health in brazil in times of coronavirus. **HU Revista**, v. 50, p. 1–8, ago. 2024. Disponível em: https://periodicos.ufjf.br/index.php/hurevista/article/view/44375.

NOGUEIRA, Thallys da Silva; SARTE, Laisa Marcorela Andreoli; GOLIATT, Priscila Vanessa Zabala Capriles. Social network x and mental health in brazil in times of coronavirus. **HU Revista**, v. 50, p. 1–8, ago. 2024. Disponível em: https://periodicos.ufjf.br/index.php/hurevista/article/view/44375.

NOGUEIRA, Thallys da Silva; SIQUEIRA, Kennya Beatriz; GOLIATT, Priscila Vanessa Zabala Capriles. Observatório do consumidor: Potencializando a indústria láctea com business intelligence. In: **Anais do Dairy Talks - Prosas de leiteiros: ciência, tecnologia, engenharia e inovação**. Juiz de Fora (MG): UFJF, 2024. Acesso em: 17/01/2025.

NOGUEIRA, Thallys da Silva. **Mineração de dados em rede social para avaliação de tendências de consumo do queijo artesanal no Brasil**. 2021. 112 p. Dissertação (Dissertação de Mestrado) — Universidade Federal de Juiz de Fora, Juiz de Fora, 2021. Disponível em: https://repositorio.ufjf.br/jspui/handle/ufjf/12524.

NOGUEIRA, Thallys da Silva; FONSECA, Leonardo Goliatt da; GOLIATT, Priscila Vanessa Zabala Capriles. Molecular dynamics modeling and simulation of tobermorite under pre-salt conditions. 2022.

NOGUEIRA, Thallys da Silva; MOURO, Vitor Agostinho; SIQUEIRA, Kennya Beatriz; GOLIATT, Priscila Vanessa Zabala Capriles. Analysis of the brazilian artisanal cheese

market from the perspective of social networks. In: ABRAHAM, Ajith; GANDHI, Niketa; HANNE, Thomas; HONG, Tzung-Pei; RIOS, Tatiane Nogueira; DING, Weiping (Ed.). **Intelligent Systems Design and Applications - 21st International Conference on Intelligent Systems Design and Applications (ISDA 2021) Held During December 13-15, 2021**. Springer, 2021. (Lecture Notes in Networks and Systems, v. 418), p. 900–909. Disponível em: https://doi.org/10.1007/978-3-030-96308-8_84.

NOGUEIRA, Thallys da Silva; SIQUEIRA, Kennya Beatriz; GOLIATT, Priscila Vanessa Zabala Capriles. Construction of a training dataset for a sentiment analysis model of dairy products tweets in brazil. **Social Network Analysis and Mining**, v. 14, n. 1, p. 85, April 2024. ISSN 1869-5469. Disponível em: https://doi.org/10.1007/s13278-024-01254-5.

NOGUEIRA, T. S.; FRANCO, A. L.; SILVA, D. H. C.; SIQUEIRA, K. B.; GOLIATT, P. V. Z. C. Mineração de dados em tweets para análise do consumo de lácteos no brasil. **The Journal of Engineering and Exact Sciences**, v. 8, n. 10, p. 14863–01a, dez. 2022. Disponível em: https://periodicos.ufv.br/jcec/article/view/14863.

NOGUEIRA, T. S.; FRANCO, A. L.; SILVA, D. H. C.; SIQUEIRA, K. B.; GOLIATT, P. V. Z. C. Quais os derivados lácteos mais consumidos no brasil durante a pandemia da covid-19? **XXVI Workshop de Iniciação Científica da Embrapa Gado de Leite - 145 PIBIC/CNPq 2021/2022**, nov. 2022. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1148821/1/Quais-os-derivados-lacteos-mais-consumidos-no-Brasil.pdf.

NOVO-LOURéS, María; RUANO-ORDáS, David; PAVóN, Reyes; LAZA, Rosalía; GóMEZ-MEIRE, Silvana; MéNDEZ, José R. Enhancing representation in the context of multiple-channel spam filtering. **Information Processing Management**, v. 59, n. 2, p. 102812, 2022. ISSN 0306-4573. Disponível em: https://www.sciencedirect.com/science/article/pii/S0306457321002879.

OLIVEIRA, Guilherme Vieira de; OLIVEIRA, Mariana de; SANTOS, Aline Gonçalves dos; MOURA, Ricardo Ribeiro; JUNIOR, Lazaro Antônio da Fonseca. Aplicação do business intelligence na gestão da cadeia de suprimentos. **Brazilian Journal of Production Engineering**, v. 9, n. 5, p. 60–69, out. 2023. Disponível em: https://periodicos.ufes.br/bjpe/article/view/42709.

OLIVEIRA, Therys Senna de Castro; NOGUEIRA, Thallys da Silva; GOLIATT, Priscila Vanessa Zabala Capriles; SIQUEIRA, Kennya Beatriz. **Observatório do Consumidor: Explorando os consumidores de lácteos no Brasil**. 2024. Vinculo, Programa de Pós-Graduação em Modelagem Computacional/Universidade Federal de Juiz de Fora; Universidade Federal de Viçosa; Empresa Brasileira de Pesquisa Agropecuária/ Embrapa Gado de Leite. Disponível em: https://cbcta2024.com.br/evento/cbcta/trabalhosaprovados/naintegra/20290.

OTWELL, Taylor. **Laravel: The PHP Framework for Web Artisans**. [S.l.], 2011. Disponível em: https://laravel.com. Acesso em: 8 out. 2024.

OUZZANI, Mourad; HAMMADY, Hossam; FEDOROWICZ, Zbys; ELMAGARMID, Ahmed. Rayyan—a web and mobile app for systematic reviews. **Systematic Reviews**, v. 5, n. 1, p. 210, 2016. ISSN 2046-4053. Disponível em: http://dx.doi.org/10.1186/s13643-016-0384-4.

PAGE, Matthew J; MCKENZIE, Joanne E; BOSSUYT, Patrick M; BOUTRON, Isabelle; HOFFMANN, Tammy C; MULROW, Cynthia D; SHAMSEER, Larissa; TETZLAFF, Jennifer M; AKL, Elie A; BRENNAN, Sue E; CHOU, Roger; GLANVILLE, Julie; GRIMSHAW, Jeremy M; HRÓBJARTSSON, Asbjørn; LALU, Manoj M; LI, Tianjing; LODER, Elizabeth W; MAYO-WILSON, Evan; MCDONALD, Steve; MCGUINNESS, Luke A; STEWART, Lesley A; THOMAS, James; TRICCO, Andrea C; WELCH, Vivian A; WHITING, Penny; MOHER, David. The prisma 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, BMJ Publishing Group Ltd, v. 372, 2021. Disponível em: https://www.bmj.com/content/372/bmj.n71.

PALMER, D. D. **Text preprocessing**: In handbook of natural language processing. 2nd edition. ed. [S.l.]: Chapman and Hall/CRC, 2010.

PANG, Bo; LEE, Lillian. **Opinion Mining and Sentiment Analysis**. [S.l.: s.n.], 2008.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEREZ, Luis; WANG, Jason. The effectiveness of data augmentation in image classification using deep learning. **ArXiv**, abs/1712.04621, 2017.

PHILLIPS, Edward M; FRATES, Elizabeth P; PARK, David J. Lifestyle medicine. **Physical Medicine and Rehabilitation Clinics of North America**, v. 31, n. 4, p. 515–526, 2020. Disponível em: https://app.dimensions.ai/details/publication/pub.1130722450.

PINTO, Cíntia Loos; VIANA, Lilian Carolina; OLIVEIRA, Solange Riveli de; SETE, Ricardo de Souza. Consumo de queijos finos: aspectos simbólicos e identitários. **Ciências Sociais Aplicadas em Revista**, v. 12, n. 22, p. 71–87, 2012.

PINTO, Herbert Laroca; ROCIO, Vitor. Combining sentiment analysis scores to improve accuracy of polarity classification in mooc posts. In: OLIVEIRA, Paulo Moura; NOVAIS, Paulo; REIS, Luís Paulo (Ed.). **Progress in Artificial Intelligence**. Cham: Springer International Publishing, 2019. p. 35–46. ISBN 978-3-030-30241-2.

PRATAMA, Enda Esyudha; RIPANTI, Eva Faja. Analysis of student academic performance and social media activities by using data mining approach. In: **Proceedings of the 2020 The 6th International Conference on E-Business and Applications**. New York, NY, USA: Association for Computing Machinery, 2020. (ICEBA 2020), p. 111–115. ISBN 9781450377355. Disponível em: https://doi.org/10.1145/3387263.3387279.

PREOţIUC-PIETRO, Daniel; VOLKOVA, Svitlana; LAMPOS, Vasileios; BACHRACH, Yoram; ALETRAS, Nikolaos. Studying user income through language, behaviour and affect in social media. **PLOS ONE**, v. 10, n. 9, p. e0138717, 2015. Disponível em: https://doi.org/10.1371/journal.pone.0138717.

PYTHON SOFTWARE FOUNDATION. **Python Language Reference. Version 3.7**. 2021. Disponível em: https://bityli.com/zMnhv.

PYTUBE. **PyTube: A lightweight, dependency-free Python library for downloading YouTube videos**. 2015. https://github.com/pytube/pytube. Accessed: 2025-01-07.

RAHMAN, S.; JAHAN, N.; SADIA, F.; MAHMUD, I;. Social crisis detection using twitter based text mining-a machine learning approach. 2023.

RANGEL, Djalma Araujo. Uma plataforma de business intelligence para análise de informações estudantis no ifpe campus igarassu. Universidade Federal da Paraíba, 2024.

RATHKE, Brandon H.; YU, Hanzhang; HUANG, Huizhong. What remains now that the fear has passed? developmental trajectory analysis of covid-19 pandemic for co-occurrences of twitter, google trends, and public health data. **Disaster Medicine and Public Health Preparedness**, v. 17, p. e471, June 2023.

R CORE TEAM. R: A language and environment for statistical computing. Vienna, Austria, 2020.

REN, Shiqi. Multi-media content clustering and computer intelligent analysis by text mining. In: **2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture**. New York, NY, USA: Association for Computing Machinery, 2022. (AIAM2021), p. 2276–2285. ISBN 9781450385046. Disponível em: https://doi.org/10.1145/3495018.3501088.

REYES-MENENDEZ, Ana; SAURA, Jose Ramon; FILIPE, Ferrão. Marketing challenges in the metoo era: gaining business insights using an exploratory sentiment analysis. **Heliyon**, v. 6, n. 3, p. e03626, 2020. ISSN 2405-8440. Disponível em: https://www.sciencedirect.com/science/article/pii/S2405844020304710.

RIPPE, James. **Lifestyle medicine.** [S.l.: s.n.], 1999.

RODRIGUES, Laura Destro; SIQUEIRA, Kennya Beatriz; NOGUEIRA, Thallys da Silva; GOLIATT, Priscila Vanessa Zabala Capriles. **Padrões de consumo de queijos no Brasil: Uma abordagem por meio do Observatório do Consumidor**. 2024. Acesso em: 17/01/2025. Disponível em: https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1169624/1/Padroes-de-consumo-de-queijos-no-Brasil.pdf.

ROTH, Dan; ZELENKO, Dmitry. Part of speech tagging using a network of linear separators. In: **Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2**. USA: Association for Computational Linguistics, 1998. (ACL '98/COLING '98), p. 1136–1142. Disponível em: https://doi.org/10.3115/980691.980755.

RUGGIERO, M. endências do consumo de lácteos em 2022 –como estamos até o momento? **Fórum Milkpoint.**, 2022.

RUSSELL, S.; NORVIG, P. Artificial intelligence - a modern approach. **Prentice-Hall**, 1995.

SAAD, Nor Hasliza Md; SAN, Chin Wei; YAACOB, Zulnaidi. Twitter sentiment analysis of the low-cost airline services after covid-19 outbreak: The case of airasia. **Business Systems Research Journal**, v. 14, n. 2, p. 1–23, 2023. Disponível em: https://doi.org/10.2478/bsrj-2023-0009.

SAGNER, M.; KATZ, D.; EGGER, G.; LIANOV, L.; SCHULZ, K.-H.; BRAMAN, M.; BEHBOD, B.; PHILLIPS, E.; DYSINGER, W.; ORNISH, D. Lifestyle medicine potential for reversing a world of chronic disease epidemics: from cell to community. **International Journal of Clinical Practice**, v. 68, n. 11, p. 1289–1292, 2014. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijcp.12509.

SANTOS, Cristina Mamédio da Costa; PIMENTA, Cibele Andrucioli de Mattos; NOBRE, Moacyr Roberto Cuce. The pico strategy for the research question construction and evidence search. **Revista Latino-Americana de Enfermagem**, Escola de Enfermagem de Ribeirão Preto / Universidade de São Paulo, v. 15, n. Rev. Latino-Am. Enfermagem, 2007 15(3), Jun 2007. ISSN 0104-1169. Disponível em: https://doi.org/10.1590/S0104-11692007000300023.

SARDINHA, Tony Berber. Lingüística de corpus: histórico e problemática. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 16, 01 2000.

SAURA, José; REYES-MENENDEZ, Ana; NOSTRAND, Chris. Does seo matter for startups? identifying insights from ugc twitter communities. **Informatics**, v. 7, p. 47, 10 2020.

SAURA, José Ramón; DEBASA, Felipe; REYES-MENENDEZ, Ana. Does user generated content characterize millennials' generation behavior? discussing the relation between sns and open innovation. In: . [S.l.: s.n.], 2019.

SAURA, Jose Ramon; PALACIOS-MARQUéS, Daniel; RIBEIRO-SORIANO, Domingo. Using data mining techniques to explore security issues in smart living environments in twitter. **Comput. Commun.**, Elsevier Science Publishers B. V., NLD, v. 179, n. C, p. 285–295, nov 2021. ISSN 0140-3664. Disponível em: https://doi.org/10.1016/j.comcom.2021.08.021.

SAURA, Jose Ramon; REYES-MENENDEZ, Ana; BENNETT, Dag R. How to extract meaningful insights from ugc: A knowledge-based method applied to education. **Applied Sciences**, v. 9, n. 21, 2019. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/9/21/4603.

SAURA, Jose Ramon; REYES-MENENDEZ, Ana; FILIPE, Ferrão. Comparing data-driven methods for extracting knowledge from user generated content. **Journal of Open Innovation: Technology, Market, and Complexity**, v. 5, n. 4, 2019. ISSN 2199-8531. Disponível em: https://www.mdpi.com/2199-8531/5/4/74.

SAURA, Jose Ramon; REYES-MENENDEZ, Ana; PALOS-SANCHEZ, Pedro. Are black friday deals worth it? mining twitter users' sentiment and behavior response. **Journal of Open Innovation: Technology, Market, and Complexity**, v. 5, n. 3, 2019. ISSN 2199-8531. Disponível em: https://www.mdpi.com/2199-8531/5/3/58.

SAURA, José Ramón; REYES-MENéNDEZ, Ana; DEMATOS, Nelson; CORREIA, Marisol B. Identifying startups business opportunities from ugc on twitter chatting: An exploratory analysis. **Journal of Theoretical and Applied Electronic Commerce Research**, v. 16, n. 6, p. 1929–1944, 2021. ISSN 0718-1876. Disponível em: https://www.mdpi.com/0718-1876/16/6/108.

SENNRICH, Rico; HADDOW, Barry; BIRCH, Alexandra. **Improving Neural Machine Translation Models with Monolingual Data**. 2016. Disponível em: https://arxiv.org/abs/1511.06709.

SERENGIL, Sefik Ilkin; OZPINAR, Alper. Hyperextended lightface: A facial attribute analysis framework. In: **2021 International Conference on Engineering and Emerging Technologies (ICEET)**. [S.l.: s.n.], 2021. p. 1–4. ISSN 2409-2983.

SILVA, DH da C; MONTEIRO, ALF; NOGUEIRA, T da S; LANA, MS; SIQUEIRA, KB; GOLIATT, PVZC; SILVA, DARLAN HENRIQUE DA COSTA; MONTEIRO, ANNA LETÍCIA FRANCO et al. Análise exploratória do interesse por marcas e produtos lácteos no brasil. 2024.

SILVA, D. H. C.; FRANCO, A. L.; NOGUEIRA, T. S.; SIQUEIRA, K. B.; GOLIATT, P. V. Z. C. Análise exploratória do interesse por lácteos orgânicos. **XXVI Workshop de Iniciação Científica da Embrapa Gado de Leite - 145 PIBIC/CNPq 2021/2022**, 2022. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1148823/1/Analise-exploratoria-do-interesse-por-lacteos-organicos.pdf.

SILVA, Priscilla de Souza. **Avaliação do desempenho de métodos de análise de sentimentos na presença das figuras de linguagem sarcasmo e ironia**. 2016. 115 p. Dissertação (Trabalho de Conclusão de Curso (Graduação)) — Universidade Federal do Sul e Sudeste do Pará, Campus Universitário de Marabá, Instituto de Geociências e Engenharias, Faculdade de Computação e Engenharia Elétrica, Curso de Bacharelado em Sistemas de Informação, Marabá, 2016. Disponível em: http://repositorio.unifesspa.edu.br/handle/123456789/233.

SIQUEIRA, KB; NOGUEIRA, TS; SOARES, ND; CAMPOS, EWAM; MORAES, EAP; VILLELA, RMMB; DAVID, JMN; GOLIATT, PVZC; DINIZ, FH; PIRES, M et al. Covid-19 e o impacto no consumo de leite e derivados. 2021.

SIQUEIRA, KB; NOGUEIRA, TS; SOARES, ND; CAMPOS, EWAM; MORAES, EAP; VILLELA, RMMB; DAVID, JMN; GOLIATT, PVZC. O impacto da pandemia no consumo de lácteos no brasil: uma abordagem das redes sociais. In: SIQUEIRA, KB (ed.). Na era do consumidor: uma visão do mercado lácteo, 2021.

SIQUEIRA, KB; NOGUEIRA, TS; SOARES, ND; CAMPOS, EW; MORAES, EAP; VILLELA, RMMB; DAVID, JMN; GOLIATT, PVZC; DINIZ, FH; PIRES, M et al. Pandemia, leite, queijo e tweets. **Anuário do Leite**, 2021.

SIQUEIRA, KB; NOGUEIRA, T da S; GOLIATT, PVZC et al. Os queijos preferidos em minas gerais. **Milkpoint**, 2024.

SIQUEIRA, K. B. **O mercado consumidor de leite e derivados**. [S.l.], 2019. 1–17 p. (Circular Técnica Embrapa, 120).

SIQUEIRA, Kennya Beatriz. **Na era do consumidor: uma visão do mercado lácteo brasileiro.** [S.l.: s.n.], 2021.

SIQUEIRA, K. B.; ARBEX, W. A.; PIRES, M. F. A.; DINIZ, F. H.; VICENTINI, N. M.; MORAES, E. A. P.; VILLELA, R. M. M. B.; DAVID, J. M. N.; GOLIATT, Priscila Vanessa Zabala Capriles. **OBSERVATÓRIO DO CONSUMIDOR**. 2022. Patente: Programa de Computador. Número do registro: BR512022002576-0, data de registro:

19/09/2022, título: "OBSERVATÓRIO DO CONSUMIDOR", Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial.

SIQUEIRA, Kennya Beatriz; GUIMARAES, Ygor Martins; NOGUEIRA, T da S; GUIMARÃES, YGOR MARTINS; THALLYS, DA SILVA NOGUEIRA. A geografia brasileira do consumo domiciliar de leite. In: SIQUEIRA, KB (ed.). Na era do consumidor: uma visão do mercado lácteo . . . , 2021.

SIQUEIRA, Kennya Beatriz; NOGUEIRA, TS; CAMPOS, EWAM; SOARES, ND; MORAES, EAP; VILLELA, RMMB; DAVID, JMN; GOLIATT, PVZC; NOGUEIRA, THALLYS SILVA; CAMPOS, EMERSON WENDELIN ALVES MOREIRA et al. O que mudou no consumo de lácteos com a pandemia? uma visão das redes sociais. In: SIQUEIRA, KB (ed.). Na era do consumidor: uma visão do mercado lácteo, 2021.

SIQUEIRA, Kennya B.; NOGUEIRA, Thallys S.; CAMPOS, Emerson W.; SOARES, Nedson D.; MORAES, Emerson A. P.; VILLELA, Regina M.M.B.; DAVID, José Maria N.; GOLIATT, Priscila V.Z.C. Análise exploratória da imagem dos lácteos em tempos de coronavírus. **INDÚSTRIA DE LATICÍNIOS**, n. 143, p. 64–66, 2020. ISSN 1678-7250.

SIQUEIRA, Kennya B.; NOGUEIRA, Thallys S.; CAMPOS, Emerson W.; SOARES, Nedson D.; MORAES, Emerson A. P.; VILLELA, Regina M.M.B.; DAVID, José Maria N.; GOLIATT, Priscila V.Z.C. O impacto da pandemia no consumo de lácteos no brasil. **INDÚSTRIA DE LATICÍNIOS**, n. 147, p. 36–38, 2020. ISSN 1678-7250.

SMATOV, Nurmaganbet; KALASHNIKOV, Ruslan; KARTBAYEV, Amandyk. Development of context-based sentiment classification for intelligent stock market prediction. **Big Data and Cognitive Computing**, v. 8, n. 6, 2024. ISSN 2504-2289. Disponível em: https://www.mdpi.com/2504-2289/8/6/51.

SOARES, Nedson; BRAGA, Regina; DAVID, José Maria; SIQUEIRA, Kennya; STRöELE, Victor; CAMPOS, Fernanda. Uma arquitetura para a recomendação de consumidores de queijo artesanal brasileiro. In: **Anais do XIV Brazilian e-Science Workshop**. Porto Alegre, RS, Brasil: SBC, 2020. p. 113–120. ISSN 2763-8774. Disponível em: https://sol.sbc.org.br/index.php/bresci/article/view/11189.

SOARES, Nedson Donato; BRAGA, Regina; DAVID, José Maria N; SIQUEIRA, Kennya Beatriz; NOGUEIRA, Thallys S; EW, CAMPOS. Redic: Recommendation of digital influencers of brazilian artisanal cheese. In: IN: BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 17., 2021, UBERLÂNDIA . . . . [S.l.], 2021.

SOARES, N. D.; CAMPOS, E.; NOGUEIRA, T.; SIQUEIRA, K. B.; DAVID, J. M. N.; BRAGA, R. Quais os tipos de queijos mais comentados no twitter. In: **Anais do Workshop de iniciação científica da Embrapa Gado de Leite**. Juiz de Fora, MG, Brasil: Embrapa Gado de Leite, 2021.

SOLKA, Jeffrey L. Text Data Mining: Theory and Methods. **Statistics Surveys**, Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada, v. 2, n. none, p. 94 – 112, 2008. Disponível em: https://doi.org/10.1214/07-SS016.

SOUZA, Lucas Alves Moreira de. Aplicação de aprendizado de máquina para predição de prioridade em gestão de incidentes. 2017.

SOUZA, Queila R.; QUANDT, Carlos O. Metodologia de análise de redes sociais. p. 31–63, 2008.

SRIDEVI, P.C.; VELMURUGAN, T. Impact of preprocessing on twitter based covid-19 vaccination text data by classification techniques. In: **2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)**. [S.l.: s.n.], 2022. p. 1126–1132.

SRIDHAR, Sashank; SANAGAVARAPU, Sowmya. Handling data imbalance in predictive maintenance for machines using smote-based oversampling. In: **2021 13th International Conference on Computational Intelligence and Communication Networks (CICN)**. [S.l.: s.n.], 2021. p. 44–49.

THAKUR, Nirmalya. Investigating and analyzing self-reporting of long covid on twitter: Findings from sentiment analysis. **Applied System Innovation**, v. 6, n. 5, 2023. ISSN 2571-5577. Disponível em: https://www.mdpi.com/2571-5577/6/5/92.

THAKUR, Nirmalya; HAN, Chia Y. **An Exploratory Study of Tweets about the SARS-CoV-2 Omicron Variant: Insights from Sentiment Analysis, Language Interpretation, Source Tracking, Type Classification, and Embedded URL Detection**. 2022. Disponível em: https://arxiv.org/abs/2208.10252.

ULFATRIYANI, Hesty; NUGROHO, Hanung Adi; SOESANTI, Indah. Implementing term frequency-inverse term frequency at tweets in indonesian fraud crime cases. In: **2020 3rd International Conference on Information and Communications Technology (ICOIACT)**. [S.l.: s.n.], 2020. p. 185–190.

ULTRALYTICS. **Multi-Modal Model - Glossary**. 2025. Acesso em: 27 jan. 2025. Disponível em: https://www.ultralytics.com/pt/glossary/multi-modal-model.

URCO, Christian Fabián Castillo; SAá, Marcelo Javier Mancheno; MURILLO, Daniel Esteban Chamorro; SALINAS, Jenny Margoth Gamboa. Felicidade no trabalho na geração dos millennials, novos desafios para os administradores / happiness at work in the millennial generation, new challenges for managers. **Brazilian Journal of Development**, v. 5, n. 9, p. 14571–14582, Sep. 2019. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/3135.

USSELMANN, Henning; AHMAD, Rangina; SIEMON, Dominik. A personality mining system for german twitter posts with global vectors word embedding. **IEEE Access**, v. 9, p. 165576–165610, 2021. ISSN 2169-3536.

UTAMI, Mailia Putri; NURHAYATI, Oky Dwi; WARSITO, Budi. Hoax information detection system using apriori algorithm and random forest algorithm in twitter. In: **2020 6th International Conference on Interactive Digital Media (ICIDM)**. [S.l.: s.n.], 2020. p. 1–5.

VERíSSIMO, Bruna; LEPRE, Larissa; TINCANI, Daniela. Diferenças entre pesquisa de marketing e pesquisa de neuromarketing. 2018.

VICENTE, J. A. G.; SIRQUEIRA, T. F. M. Uma ferramenta de análise estática de código fonte para aplicações web php. **Caderno de Estudos em Sistemas de Informação**, v. 7, n. 1, 2022.

VIEIRA, R.; LIMA, V. L. S. Lingüística computacional: princípios e aplicações. In: NEDEL, Luciana (Ed.). **IX Escola de Informática da SBC-Sul**. [S.l.]: SBC-Sul, 2001.

WITTEN, Ian H.; FRANK, Eibe. Data mining pratical machine learning tools and techniques. Elsevier, 2005.

ZHANG, J.; MANI, I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In: **Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets**. [S.l.: s.n.], 2003.

ZHUANG, Fuzhen; QI, Zhiyuan; DUAN, Keyu; XI, Dongbo; ZHU, Yongchun; ZHU, Hengshu; XIONG, Hui; HE, Qing. **A Comprehensive Survey on Transfer Learning**. 2020. Disponível em: https://arxiv.org/abs/1911.02685.

ZOU, Boyuan; PENG, Bo; HUANG, Qunying. Flood depth assessment with location-based social network data and google street view - a case study with buildings as reference objects. In: **IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium**. [S.l.: s.n.], 2022. p. 1344–1347. ISSN 2153-7003.