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

Bruno Henrique Rodrigues

Análise de consumo de produtos à base de leitelho através de técnicas de mineração de dados

Juiz de Fora

2025

Bruno Henrique Rodrigues

Análise de consumo de produtos à base de leitelho através de técnicas de mineração de dados

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

Orientadora: Prof[a]. Dr[a]. Priscila Vanessa Zabala Capriles Goliatt

Juiz de Fora

2025

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Rodrigues, Bruno Henrique.
Análise de consumo de produtos à base de leitelho através de técnicas de mineração de dados / Bruno Henrique Rodrigues. -- 2025.
64 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2025.

1. Laticínios. 2. Análise mercadológica. 3. Aprendizado de máquina. 4. Análise de sentimentos. 5. Reconhecimento facial. I. Goliatt, Priscila Vanessa Zabala Capriles, orient. II. Título.

Bruno Henrique Rodrigues

Análise do consumo de produtos à base de leitelho através de técnicas de mineração de dados

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

Aprovada em 25 de agosto de 2025.

BANCA EXAMINADORA

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

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

Prof.ª Dr.ª Ana Clarissa dos Santos Pires
Universidade Federal de Viçosa

Juiz de Fora, 24/08/2025.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 30/08/2025, às 16:29, 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 **Ana Clarissa dos Santos Pires**, **Usuário Externo**, em 01/09/2025, às 14:48, 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 **Heder Soares Bernardino**, **Professor(a)**, em 03/09/2025, às 13:38, 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 **2573460** e o código CRC **9716066C**.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Universidade Federal de Juiz de Fora (UFJF), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG).

RESUMO

O leitelho é um coproduto originado na produção da manteiga, e sua composição é semelhante à do leite desnatado. Apesar da alta produção de manteiga no Brasil e dos valores nutricionais e industriais que o leitelho possui, esse coproduto não é encontrado como ingrediente com tanta facilidade, além de ser altamente poluente quando descartado de forma incorreta. Assim, buscamos comparar a tendência e padrão de consumo de produtos à base de leitelho por pessoas dentro e fora do Brasil. Para isso, utilizamos os resultados do formulário "Estudo do comportamento brasileiro sobre o consumo de produtos lácteos" e também coletamos, com raspagem de dados, detalhes de produtos com leitelho em lojas *online* e suas avaliações junto com informações básicas dos avaliadores, imagem de perfil e país. As opiniões obtidas nas duas fontes de dados foram categorizadas, com análise de sentimentos do conteúdo textual, por polaridade (positiva, negativa ou neutra), e foram determinados sexo e faixa etária processando as imagens de perfil válidas coletadas na raspagem de dados. Dentre os respondentes do formulário, 58% não eram nada familiares com o termo "leitelho", dos quais 40% deles responderam que os motivos para não experimentar seriam produtos à base dele por ter "sabor ruim" ou "textura ruim". "Presença de substâncias não adequadas à saúde humana" e "qualidade nutricional inferior" também foram respostas frequentes, representando 39% do total, apesar de outros estudos já mostrarem o contrário. Foi possível observar que, apesar dos benefícios do leitelho, há pouca variedade de produtos industrializados que o utilizam na composição, foram identificados pouco conhecimento e também pouco consumo de leitelho e seus produtos pelas respostas do formulário. No entanto, foram poucas respostas negativas para "palavra, sentimento e/ou emoção", e os principais motivos para experimentar produtos à base dele são, principalmente, sabor e curiosidade. Junto a isso, as categorias de produtos que os participantes têm mais interesse em experimentar, como panificados, estão alinhadas com as categorias dos produtos internacionais coletados. A distribuição de faixa etária estimada para esses avaliadores foi similar à dos participantes do formulário, majoritariamente entre 25 e 44 anos, e as avaliações coletadas também foram majoritariamente classificadas com sentimentos positivos, mostrando que existe um mercado a ser explorado, trazendo benefícios para o produtor, para o consumidor, e também como uma forma de reduzir possíveis descartes inadequados. Cabe a trabalhos futuros aprimorar a raspagem de dados, como aumentar a frequência das raspagens ou coletar mais informações dos usuários que avaliaram as compras para mapeamento de perfil de compra.

Palavras-chave: Laticínios. Análise mercadológica. Aprendizado de máquina. Análise de Sentimentos. Reconhecimento facial.

ABSTRACT

Buttermilk is a by-product of butter production, and its composition is similar to that of skim milk. Despite the high production of butter in Brazil and the nutritional and industrial values that buttermilk has, this co-product is not easily found as an ingredient, and its disposal is highly polluting when done incorrectly. Thus, we sought to compare the trends and consumption patterns of buttermilk-based products by people inside and outside Brazil. For this, we used the results of the form "Estudo do comportamento brasileiro sobre o consumo de produtos lácteos" and we also collected, with data scraping, details of products with buttermilk in online stores and their evaluations, along with basic information of the evaluators, profile picture, and country. The opinions obtained in the two data sources were categorized, with sentiment analysis of the textual content, by polarity (positive, negative, or neutral), and sex and age group were determined by processing the valid profile images collected in the data scraping. Among the respondents, 58% were not familiar with the term “buttermilk”, of which 40% answered that they would not consume products based on it because it would have “bad taste” or “bad texture”. “Presence of substances not suitable for human health” and “inferior nutritional quality” were also frequent responses, representing 39% of the total, despite other studies already showing the opposite. It was possible to observe that, despite the benefits of buttermilk, very few variety of industrialized products use it in the composition, little knowledge and also little consumption of buttermilk and its products were identified by the responses to the form. However, there were few negative answers for “word, feeling and/or emotion” about buttermilk, and the main reasons for trying buttermilk products are mainly flavor and curiosity. In addition, the categories of products that participants are most interested in trying, such as baked goods, are aligned with the categories of international products collected. The estimated age distribution for users who evaluated buttermilk-based products was similar to that of the participants in the form, mostly between 25 and 44 years old. The evaluations collected were also mostly classified with positive feelings, showing that there is a market to be explored, bringing benefits to the producer, to consumers, and also as a way to reduce possible inappropriate discards. Future studies should improve data scraping, such as increasing the frequency of scraping or collecting more information from users who evaluated purchases for purchase profile mapping.

Keywords: dairy; market analysis; machine learning; sentiment analysis; facial recognition.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O Brasil é o terceiro maior produtor de leite do mundo, sendo produzidos cerca de 35 bilhões de litros por ano (Brasil, 2025). Além do leite, são produzidas e consumidas toneladas de seus derivados, como queijo, iogurte e manteiga. No ano de 2024, foram fabricadas 775 toneladas de queijo, e 83 mil toneladas de manteiga no país, sendo o 11º maior produtor desse derivado, representando 0,7% da produção mundial, cerca de 2% a mais que no ano anterior, e ainda é previsto que haja um aumento anual de 3,6% em todo o mercado mundial até 2027 (MilkPoint, 2023).

No processo de produção da manteiga, que envolve a bateção do creme refrigerado, ocorre a aglomeração de glóbulos de gordura e separação da fase líquida, na qual essa aglomeração é lavada e se torna a manteiga, enquanto a fase líquida, produzida na mesma proporção, geralmente não é convertida em produtos alimentícios, especialmente, em laticínios de médio e pequeno porte. O leitelho é esse coproduto líquido, com propriedades nutricionais semelhantes às do leite desnatado, com mais fosfolipídeos e membrana do glóbulo de gordura do leite (Machado; Ramos; Antunes, 2022), podendo ser usado como seu substituto na indústria alimentícia. Porém, essa utilização do leitelho não acontece de forma industrial, diferente do soro de leite, o qual é coproduzido de 8 a 9 litros para cada quilo de queijo fabricado (Oliveira; Bravo; Tonial, 2012), e seus produtos.

Para estudar esses possíveis pontos de oportunidade, analisamos as variedades de produtos com leitelho na composição, nacionais e internacionais, as avaliações de compras. Também comparamos as opiniões e níveis de conhecimento que as pessoas têm sobre esse coproduto, além de comparar os produtos que elas teriam interesse em experimentar com o que existe à venda. Com esses resultados, visamos levantar o potencial mercadológico com os benefícios, para a indústria e para o consumidor, da utilização e do consumo do leitelho.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Analisar o consumo de leitelho e produtos à base dele, a partir de raspagem de dados de produtos vendidos *online* e do formulário "Estudo do comportamento brasileiro sobre o consumo de produtos lácteos", fornecido pelo grupo de pesquisa Termodinâmica Molecular Aplicada (THERMA) da Universidade Federal de Viçosa (UFV), para identificar possíveis pontos de oportunidade para sua utilização na produção de alimentos em nível industrial.

2.2 OBJETIVOS ESPECÍFICOS

- Coletar dados de produtos que contêm leitelho em sua composição, suas avaliações e informações dos usuários que os avaliaram;
- Realizar análise exploratória dos dados do formulário e da raspagem de dados (*web scraping*);
- Comparar as análises de sentimentos das palavras respondidas nas opiniões do formulário com as avaliações coletadas dos usuários;
- Analisar variedades dos produtos e perfil de seus compradores, nacionais e internacionais.

3 REFERENCIAL TEÓRICO

3.1 LEITELHO E SUAS FUNCIONALIDADES

O leitelho, conhecido em inglês como *buttermilk*, é um coproduto do leite originado na produção da manteiga e na mesma proporção que ela, com sua composição semelhante à do leite desnatado, mas com maior concentração de fosfolipídios e proteínas provenientes da membrana do glóbulo de gordura do leite, ou MGGL (Teixeira *et al.*, 2020). Apesar de seus bons valores nutricionais e industriais, ele ainda é subaproveitado no mercado nacional, além de provocar danos ao meio ambiente quando descartado sem tratamento, devido à sua alta demanda bioquímica de oxigênio (Ramos *et al.*, 2021).

O consumo de leitelho traz benefícios para a saúde, visto que esses fosfolipídeos e proteínas da MGGL têm propriedades que auxiliam na saúde cardiovascular, na resposta imunológica e também apresentaram ação anticarcinogênica em testes *in vitro* (Spitsberg, 2005). Além dos benefícios para a saúde que esses fosfolipídeos trazem, eles também possuem propriedades químicas que os tornam bons agentes emulsificantes e agentes espumantes (Dewettinck *et al.*, 2008), sendo vantajosos para a indústria alimentícia.

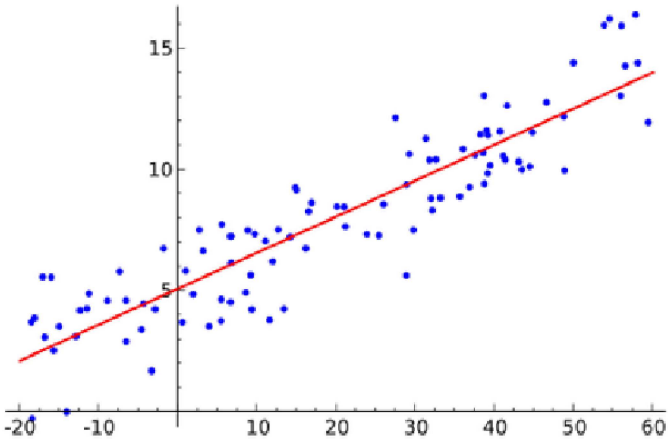
Os trabalhos de (Ramos *et al.*; Santos) são alguns dos que propõem alternativas para a utilização do leitelho, como na produção de sorvete e bebida láctea, respectivamente, mas, mesmo assim, é difícil encontrar esses e outros produtos em escala industrial, como em pães, molhos, ou misturas para panquecas, podendo estar limitado também por questões culturais de tais alimentos, ou de armazenamento e transporte (nos casos de compras *online*).

3.2 APRENDIZADO DE MÁQUINA

O aprendizado de máquina (do inglês *machine learning*) é uma área de inteligência artificial (IA) de desenvolver algoritmos de forma que eles aprendam e evoluam com base nos seus dados e seus resultados (Faceli, 2021). Duas das principais abordagens desse método são os aprendizados supervisionado e não-supervisionado (Géron, 2019).

O aprendizado de máquina supervisionado é definido por haver o resultado dos dados de entrada (variáveis respostas) já informados, ou seja, os dados são rotulados, e assim seu objetivo é predizer novas respostas com base nas anteriores já conhecidas. Quando os valores da predição são contínuos, o método é caracterizado como regressão, e quando os valores são discretos é caracterizado como classificação (Izbicki, 2020). Para a regressão, calcula-se uma equação que melhor se ajusta aos valores observados, como o modelo de regressão linear ilustrado na Figura 1, e os novos valores serão definidos pela equação calculada (Nasteski, 2017).

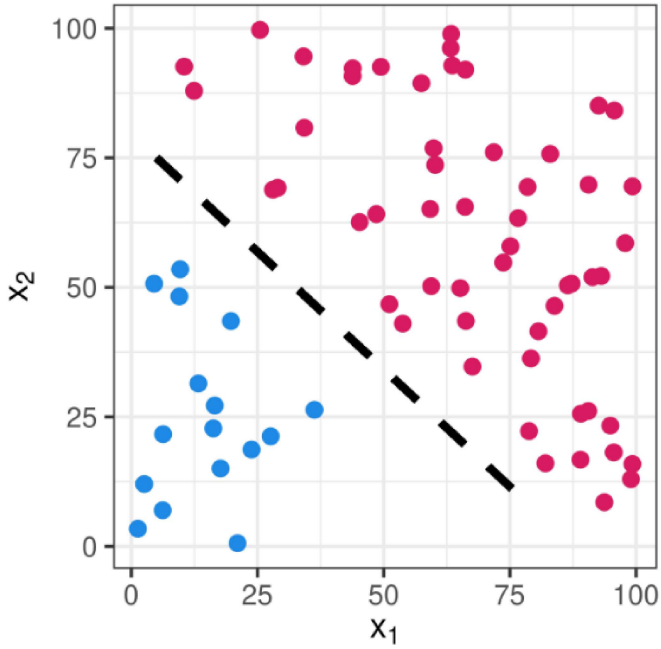
Figura 1 – Exemplo de regressão linear.



Os pontos azuis são as observações já conhecidas, enquanto a reta vermelha é uma equação linear que melhor se ajusta às observações. Fonte: Nasteski (2017).

De forma similar, o modelo de classificação envolve determinar ao menos uma hipersuperfície a partir dos dados observados, para que as novas observações sejam classificadas de acordo com suas posições referentes a esse(s) hiperplano(s), como ilustra a Figura 2 (Izbicki, 2020).

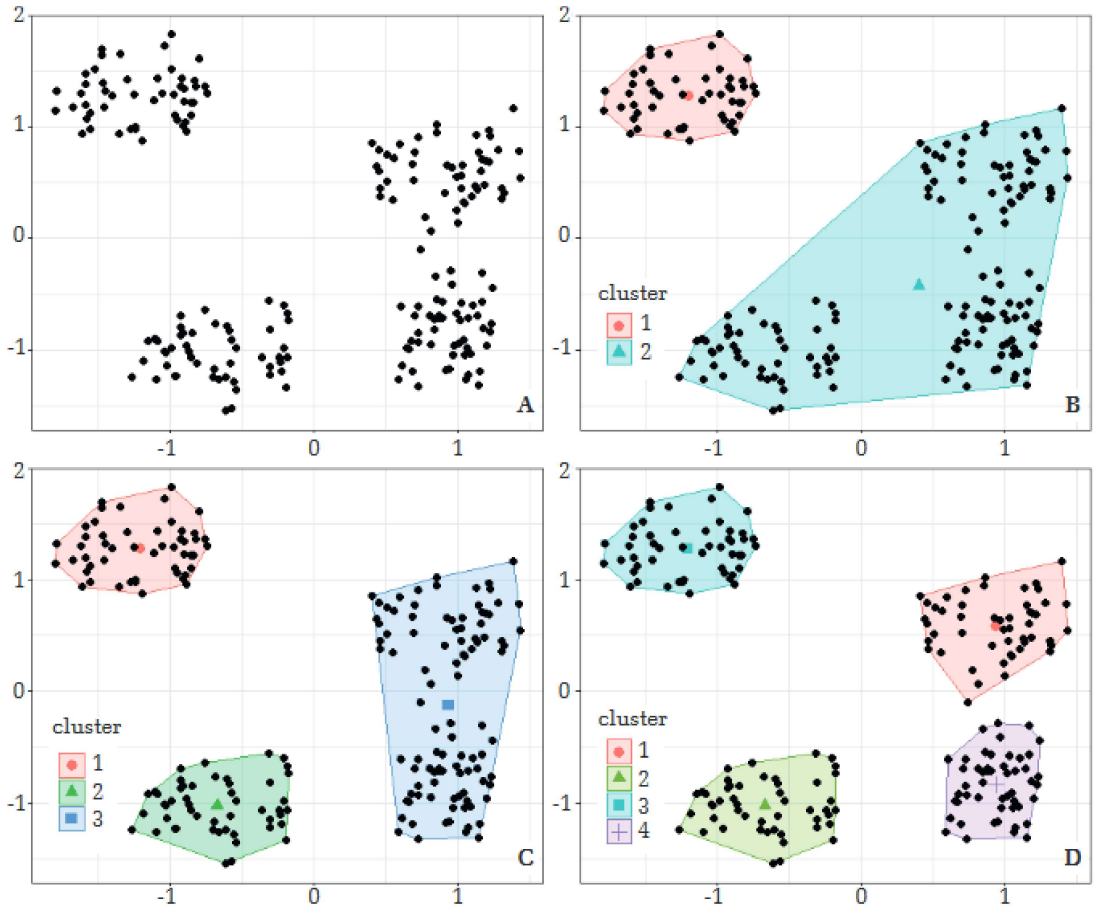
Figura 2 – Exemplo de classificação.



Os pontos são as observações já conhecidas, enquanto a reta tracejada é o hiperplano calculado que aqui define suas diferentes categorias, representadas pelas cores azul e vermelha. Fonte: Izbicki (2020).

No aprendizado não-supervisionado, não há definição de variável resposta, ou seja, os dados não possuem rótulos previamente atribuídos. Nessa abordagem, busca-se identificar padrões nos dados, o que pode ser feito por meio de técnicas de agrupamento (Alpaydın, 2020). O objetivo do agrupamento é reunir registros com características semelhantes em grupos (ou *clusters*), de modo que a distância entre os elementos de um mesmo grupo seja mínima, e a distância entre grupos distintos seja máxima. Um exemplo clássico é o algoritmo *k-means*, que forma os grupos com base na minimização da soma das distâncias quadráticas euclidianas entre os pontos e os respectivos centróides (Fernandes; Chiavegatto Filho, 2019), como ilustra a Figura 3.

Figura 3 – Exemplo de agrupamento por *k-means*.

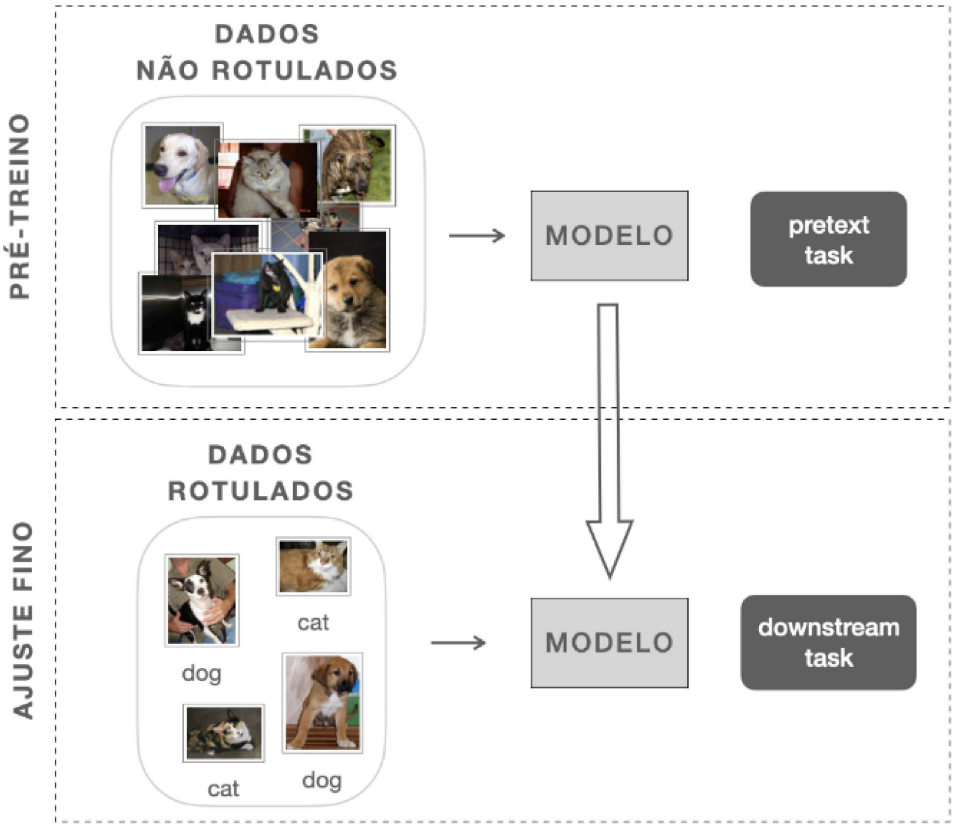


Os pontos são as observações já conhecidas. (A) Dados sem agrupamento. (B) Dados agrupados em $k = 2$ *clusters*, representados por vermelho e azul. (C) Dados agrupados em $k = 3$ *clusters*, representados por vermelho, verde e azul. (D) Dados agrupados em $k = 4$ *clusters*, representados por vermelho, verde, azul e roxo. Os pontos médios de cada *cluster* são representados por círculo no *cluster* 1, triângulo no *cluster* 2, quadrado no *cluster* 3, cruz no *cluster* 4. Fonte: Fernandes e Chiavegatto Filho (2019).

Dentre outras vertentes e variações, existe também o aprendizado auto-supervisionado, que utiliza dados não rotulados para gerar um modelo inicial e, então, utiliza dados com

rótulos para aprimorar esse modelo. Esse aprendizado é utilizado principalmente para trabalhos com imagens, áudios ou textos (Raina *et al.*, 2007; Valois, 2022). Com os dados não rotulados, é realizado um pré-treino do modelo, podendo ser ajustado com uma *pretext task*, visando reforçar o aprendizado das características dos dados de entrada, como utilizar uma mesma imagem com rotação diferente, seguido pelo ajuste fino, onde o modelo é ajustado com o uso de dados rotulados, também podendo passar por mais uma tarefa de ajuste, *downstream task*, que é a utilização desse modelo ajustado em outro conjunto de dados (Pereira, A. L., 2023), como resume a Figura 4.

Figura 4 – Exemplo do processo de aprendizado auto-supervisionado.



O modelo inicial (pré-treino) é calculado utilizando dados não rotulados, podendo ser aprimorado com *pretext task*, seguindo para o ajuste fino, agora com dados rotulados, novamente podendo ser aprimorado com *downstream task*. Fonte: Amanda Lucas Pereira (2023).

3.3 ANÁLISE DE CONTEÚDO DE DADOS TEXTUAIS

A análise de conteúdo de dados textuais, ou apenas análise de conteúdo, é o procedimento de analisar todo tipo de comunicação, verbal ou não-verbal, de simples sinais a textos completos, a fim de extrair informações de seu conteúdo (Bardin, 2011), podendo ser quantitativa e/ou qualitativa. Esse procedimento é utilizado em várias áreas, como

nas ciências sociais (Cappelle; Melo; Gonçalves, 2003; Cardoso; Oliveira; Ghelli, 2021) ou na saúde (Campos, 2004).

A análise quantitativa tem como objetivo a classificação estatística dos dados textuais a partir da quantificação deles, enquanto a qualitativa busca interpretar os sentimentos e compreender as motivações dos dados textuais (Cappelle; Melo; Gonçalves, 2003). Essas duas abordagens se complementam e podem ser trabalhadas de forma computacional com o processamento de linguagem natural, método que permite que essas análises sejam feitas de forma mais prática, principalmente quando se trata de grandes conjuntos de dados.

3.3.1 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é a transformação da língua natural humana, em dados que podem ser interpretados pelo computador, e é uma das técnicas que possibilitam o funcionamento de corretores ortográficos ou gramaticais, traduções, e análise de sentimentos (Barbosa *et al.*, 2017). Para o PLN tradicional, são necessárias algumas etapas de pré-processamento, como tokenização, remoção de *stopwords*, e *stemming* (Rodríguez; Bezerra, 2020).

O processo de tokenização é a transformação do texto em *tokens* a partir de algum separador, o mais usual sendo o espaço, mas também podendo ser um sinal de pontuação qualquer, onde cada item separado será um *token*. Também é aplicada a normalização nos *tokens*, para troca de caracteres especiais e minusculização da frase (Palmer, 2010; Nogueira; Siqueira; Capriles, 2024). A frase abaixo ilustra um exemplo de tokenização e normalização:

"O leitelho é muito bom! Já provei 3 produtos diferentes."

Aplicando o processo, a frase se torna o seguinte conjunto de *tokens*:

["o", "leitelho", "e", "muito", "bom", "!", "ja", "provei", "3", "produtos", "diferentes", "."]

As *stopwords* (palavras de parada) são as palavras que não possuem valor semântico, como artigos, preposições, pontuações ou numerais (Nogueira; Siqueira; Capriles, 2024). Seguindo com o exemplo anterior, as *stopwords* são ["o", "e", "!", "ja", "3", "."], restando apenas os *tokens* ["leitelho", "muito", "bom", "provei", "produtos", "diferentes"].

Após a remoção das *stopwords*, o processo de *stemming* é encontrar o núcleo (*stem*) da palavra, eliminando flexões, sufixos e prefixos, o que reduz o custo computacional do PLN (Rodríguez; Bezerra, 2020). No mesmo exemplo, o *stemming* é aplicado, e os *tokens* agora se tornam os seguintes:

["leit", "muit", "bom", "prov", "produt", "diferent"]

3.3.2 ANÁLISE DE SENTIMENTOS

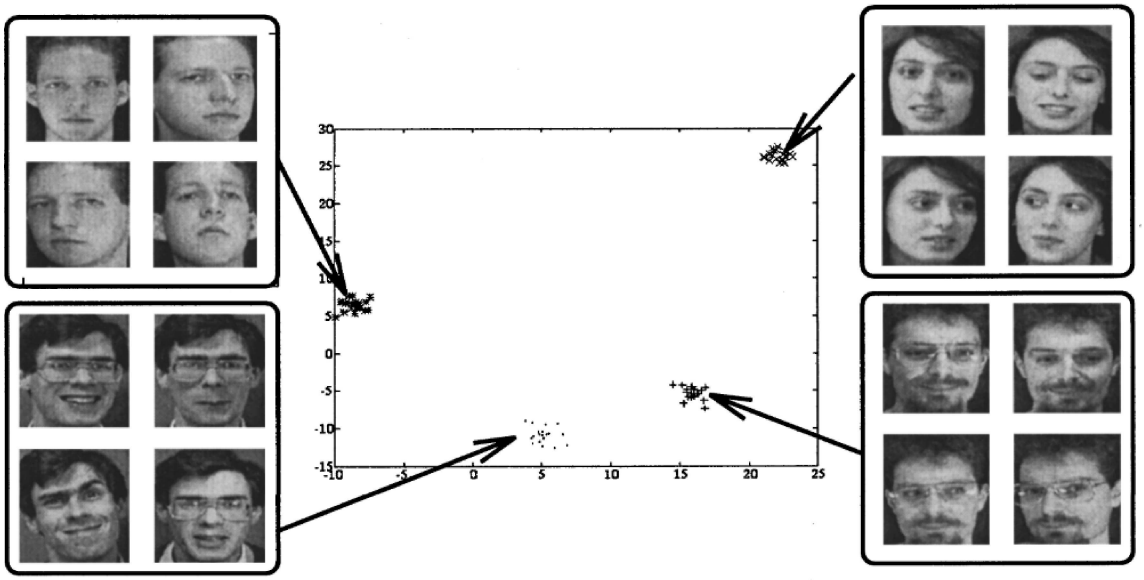
A análise de sentimentos é a análise de uma frase ou texto, após o PLN, para identificar o sentimento que aquele texto transmite, como o nome sugere, que pode ser usado para várias finalidades, como auxiliar no desenvolvimento de campanhas de *marketing* (Nogueira; Siqueira; Capriles, 2024). Essa análise também pode ser aplicada para classificação da polaridade do conteúdo textual, como positiva, negativa, ou neutra.

Considerando novamente os *tokens* do exemplo "O leitelho é muito bom! Já provei 3 produtos diferentes.", o *token* "bom" pode ser avaliado individualmente como positivo, ou até unido com "muit" para intensificar esse sentimento, e ainda que o *token* "diferent" seja de um adjetivo neutro, a frase no geral é classificada como positiva.

3.4 RECONHECIMENTO FACIAL

O reconhecimento facial é um conjunto de técnicas computacionais, como processamento de imagem e reconhecimento de padrões, usadas para detectar, identificar faces em uma imagem ou vídeo (Oliveira, 2018). A Figura 5 apresenta diferentes imagens de rostos e o agrupamento dos indivíduos por meio do aprendizado não-supervisionado (Etemad; Chellappa, 1997). Essa abordagem pode ser usada para análises estatísticas ou até verificação de identidade em sistemas de segurança, através de técnicas de aprendizado de máquina (Silva, 2018).

Figura 5 – Exemplo de agrupamentos de imagens de rostos diferentes.



Quatro imagens de rosto para cada uma das quatro pessoas diferentes são aproximadas por meio das suas características faciais. Fonte: Etemad e Chellappa (1997).

Indo além de apenas identificar a presença de rostos, também existem técnicas para reconhecimento e extração de características faciais, como sexo, raça, idade, e até emoção, como mostra a Figura 6.

Figura 6 – Exemplo de análise de atributos faciais.

Para as diferentes imagens do mesmo rosto, com diferentes expressões, são estimados idade, emoção, gênero, raça. Fonte: Taigman *et al.* (2014).

4 MATERIAL E MÉTODOS

4.1 FORMULÁRIO “ESTUDO DO COMPORTAMENTO DE BRASILEIROS SOBRE O CONSUMO DE PRODUTOS LÁCTEOS”

Os dados são provenientes de 887 respostas das diferentes regiões do Brasil, representando proporcionalmente cada região, de 17 de abril de 2024 a 10 de julho de 2024, do formulário “Estudo do comportamento de brasileiros sobre o consumo de produtos lácteos”, elaborado e compartilhado pelo grupo de pesquisa Termodinâmica Molecular Aplicada (THERMA) da Universidade Federal de Viçosa (UFV) com aprovação do comitê de ética em pequisa, com o objetivo de coletar informações a respeito da influência da neofobia alimentar sobre a percepção e o consumo de coprodutos lácteos pela população brasileira, que faz parte do trabalho “Nanociência e inteligência artificial aplicadas à cadeia de produtos lácteos: Síntese e extração de estruturas supramoleculares inovadoras estratégicas às indústrias alimentícia, farmacêutica, cosmética e de química fina”.

O formulário possui 39 perguntas, mostradas no Quadro 1, sendo dez sobre neofobia alimentar, quatro sobre o consumo de produtos lácteos, dez sobre soro de leite (com ilustração do coproduto), dez sobre leitelho (também com ilustração do coproduto), e cinco de perfil do participante. Para este trabalho, não analisamos as perguntas sobre o soro de leite, focando nas perguntas sobre o leitelho.

Quadro 1 – Perguntas do formulário organizadas por categoria.

Categoria	**Perguntas**
Neofobia alimentar	“1. Eu estou constantemente experimentando alimentos novos e diferentes.”[1], “2. Eu não confio em novos alimentos.”[1], “3. Se eu não sei o que contém um alimento, eu não experimento.”[1], “4. Eu gosto de comidas de diferentes países.”[1], “5. Comidas de outros países parecem muito estranhas para serem consumidas.”[1], “6. Em eventos sociais, eu experimento novos alimentos.”[1], “7. Eu tenho receio de comer alimentos que eu nunca experimentei antes.”[1], “8. Eu sou muito exigente em relação aos alimentos que eu escolho para comer.”[1], “9. Eu como praticamente de tudo.”[1], “10. Eu gosto de experimentar novos restaurantes de comidas de outros países.”[1]
Produtos lácteos	“11. Qual a frequência que você consome produtos lácteos?”[1], “12. Você tem o costume de ler a lista de ingredientes ao comprar produtos lácteos?”[1], “13. Quais os principais fatores você leva em consideração ao adquirir produtos lácteos?”[1], “14. Considero que a produção de leite e derivados gera um impacto negativo ao meio ambiente.”[1]
Soro de leite	“15. Quão familiar é o termo “soro de leite” para você?”[1], “16. Com que frequência você consome alimentos contendo soro como ingrediente?”[1], “17. Você teria vontade de experimentar alimentos contendo soro de leite?”[1], “18. Quais os principais motivos levariam você a experimentar produtos contendo soro de leite?”[2], “19. Quais as principais razões te desmotivariam a experimentar produtos contendo soro de leite?”[2], “20. Quais produtos contendo soro você gostaria de experimentar?”[2], “21. Primeira palavra, sentimento e/ou impressão”[3], “22. Segunda palavra, sentimento e/ou impressão”[3], “23. Terceira palavra, sentimento e/ou impressão”[3], “24. Quarta palavra, sentimento e/ou impressão”[3]
Leitelho	“25. Quão familiar é o termo “leitelho” para você?”[2], “26. Com que frequência você consome alimentos contendo leitelho como ingrediente?”[2], “27. Você teria vontade de experimentar alimentos contendo leitelho?”[2], “28. Quais os principais motivos levariam você a experimentar produtos contendo leitelho?”[2], “29. Quais as principais razões te desmotivariam a experimentar produtos contendo leitelho?”[2], “30. Quais produtos contendo leitelho você gostaria de experimentar?”[2], “31. Primeira palavra, sentimento e/ou impressão”[3], “32. Segunda palavra, sentimento e/ou impressão”[3], “33. Terceira palavra, sentimento e/ou impressão”[3], “34. Quarta palavra, sentimento e/ou impressão”[3]
Perfil do usuário	“35. Qual o seu principal meio de informação sobre os alimentos?”[2], “36. Com qual gênero você se identifica?”[3], “37. Qual a sua faixa etária?”[1], “38. Qual a sua renda?”[1], “39. Em qual estado do país você mora atualmente?”[1]

Perguntas com múltiplas alternativas para marcar estão acompanhadas por “1”, perguntas com múltiplas alternativas para marcar e também para escrever estão acompanhadas por “2”, perguntas apenas para responder apenas escrevendo estão acompanhadas por “3”.
Fonte: Elaborado pelo Autor (2025).

As respostas das perguntas “31. Primeira palavra, sentimento e/ou impressão”, “32. Segunda palavra, sentimento e/ou impressão”, “33. Terceira palavra, sentimento e/ou impressão” e “34. Quarta palavra, sentimento e/ou impressão” tiveram os sinais de pontuação removidos, por haver alguns registros como “—” ou “???”, em seguida, foram excluídos os registros que tinham tamanho menor ou igual a 1, pela impossibilidade de extrair qualquer informação dessas respostas. Para a pergunta sobre identidade de gênero, foi necessário alterar algumas respostas, como “sou mulher” ou “sou homem”, para que ficassem como as opções de marcar, “Feminino” ou “Masculino”. Apenas três pessoas não se identificaram como masculino ou feminino, sendo agrupadas em “Outros”.

Também agrupamos as respostas da pergunta “36. Com qual gênero você se identifica?”, criando uma terceira opção “Outros” para agrupar os três registros afetados, mas posteriormente foram excluídos esses registros pela pequena parcela que eles representavam no total de respostas. Outras respostas de perguntas com opções tanto abertas quanto fechadas também foram trabalhadas para que fossem agrupadas “Outros”, dadas as várias ocorrências de registros únicos. Os filtros e novas categorias foram aplicados antes dos testes estatísticos e gráficos que serão apresentados.

Para a análise de sentimentos, utilizamos uma lista de adjetivos e suas polaridades (positivo, negativo, e neutro) de (Nogueira, 2025) e também a complementando com adjetivos coletados, como “cremoso(a)” e “maravilhoso(a)”. Para essa análise, foi feita uma função para tokenizar as respostas, buscando-as nessa lista de adjetivos e, então, atribuindo um valor para a polaridade (+1 se positiva, 0 se neutra, -1 se negativa), e um contador de palavras válidas (adjetivos encontrados na lista) na resposta. Existindo pelo menos uma palavra válida, era então retornada a média da polaridade dessas palavras, sendo positiva para média $\geq +1/3$, negativa se a média $\leq$ -1/3, ou neutra caso a média estivesse entre esses dois valores.

4.2 RASPAGEM DE DADOS (*WEB SCRAPING*)

Os dados foram coletados através da raspagem de dados nos sites Amazon[1] e iHerb[2], que foram lojas que apresentaram mais diversidade de produtos à base de leitelho, além de exportarem para diversos países. Também foi possível coletar informações básicas (públicas) de perfil dos usuários que avaliaram os produtos comprados nesses sites, sendo o país e a imagem de perfil. Apesar de a Amazon apresentar produtos contendo leitelho, eram muito poucos (menos de 5 entre os 437 resultados do momento) , principalmente quando comparados com a quantidade existente no iHerb, os produtos mais recorrentes na Amazon eram livros de receitas e músicas com a palavra “*buttermilk*” no Amazon Music.

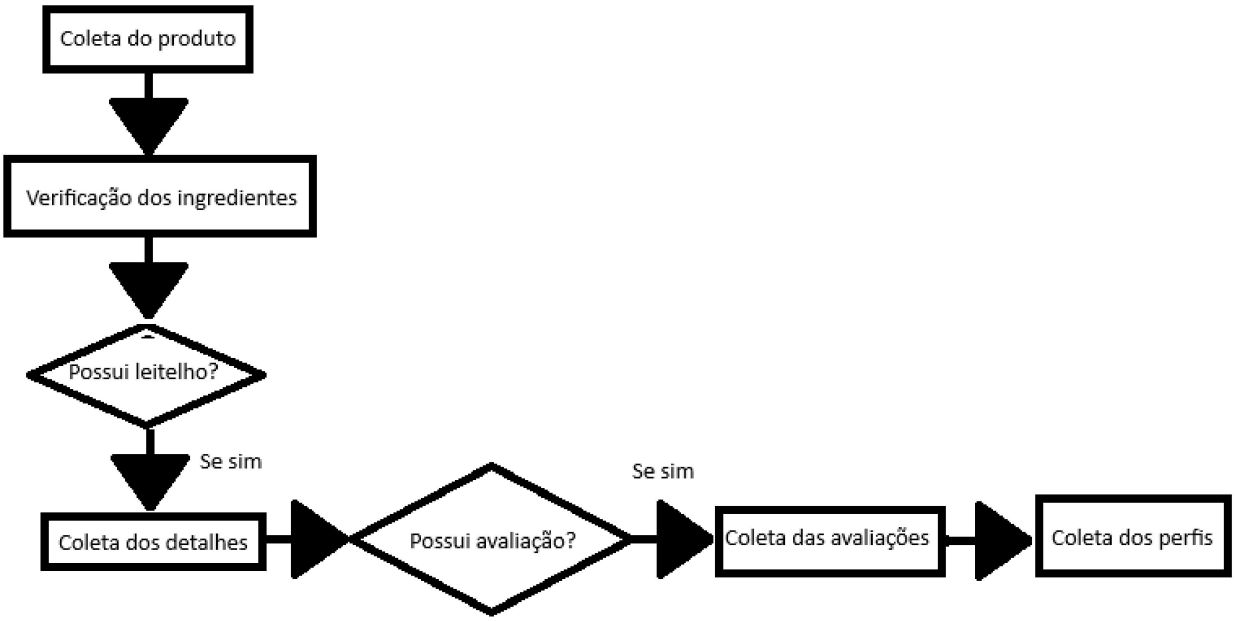
Na raspagem de dados, foi necessário avaliar os detalhes dos produtos das páginas

[1] https://www.amazon.com.br
[2] https://iherb.com

de busca, pois esses poderiam se referir aos termos “leite”, “butter” ou “milk” no nome do produto ou na lista de ingredientes, ou ainda casos de “buttermilk” traduzido pela própria loja como “soro de leite coalhado”. Caso o produto coletado contenha leitelho na sua composição, os detalhes dele são armazenados, e inicia-se o processo de coleta das avaliações escritas pelos usuários, junto com título, nome do usuário, imagem de perfil, data e país. A Figura 7 ilustra o fluxograma da raspagem dos dados.

Caso o perfil tenha uma imagem, é iniciado o processo para tentar identificar um rosto na imagem, estimando o sexo e a idade, posteriormente sendo categorizados nas faixas etárias “18 a 24 anos”, “25 a 34 anos”, “35 a 44 anos”, “45 a 54 anos”, “55 a 64 anos”, “65 anos ou mais”, seguindo as mesmas opções usadas na aplicação do formulário. Além das notas nas avaliações, também realizamos a análise de sentimentos de todas essas avaliações coletadas, já traduzidas em português pelo próprio iHerb.

Figura 7 – Fluxograma da raspagem de dados.



O algoritmo inicia buscando pela palavra desejada (“leitelho” ou “*buttermilk*”), coleta os resultados da busca, verifica a página de cada produto tentando buscar por leitelho na composição e, se, possuir avaliações, também as coleta, junto com as informações do perfil de quem avaliou. Caso não seja encontrado leitelho na descrição do produto, nenhuma avaliação ou informação de quem avaliou é coletada. Fonte: Elaborada pelo Autor (2025).

4.3 TESTES ESTATÍSTICOS

O teste Z de proporção foi aplicado para verificar se havia diferença nas proporções de respostas por gênero, por renda, ou por faixa etária. O teste Qui-quadrado de independência foi aplicado nas tabelas de contingência n-dimensionais para verificar se existia diferença estatística significativa entre as variáveis, como "gênero"-"faixa etária"-"frequência de consumo de produtos contendo leitelho". O nível de significância considerado em todos os testes foi $\alpha = 5\%$.

As hipóteses testadas foram que as variáveis de interesse estavam mais associadas a uma faixa etária e também a uma renda (no caso dos dados do formulário, dado que essa informação não foi possível de ser obtida com a raspagem de dados). Também foi incluído o gênero (para os dados do formulário) ou o sexo estimado (nos dados da raspagem de dados), testando se a distribuição de respostas segue a mesma proporção esperada entre gêneros/sexos com faixa etária ou renda, como a distribuição do nível de conhecimento do termo "leitelho" por gênero em cada renda, ou de categorias avaliadas nas avaliações coletadas por sexo estimado em cada faixa etária.

4.4 FERRAMENTAS

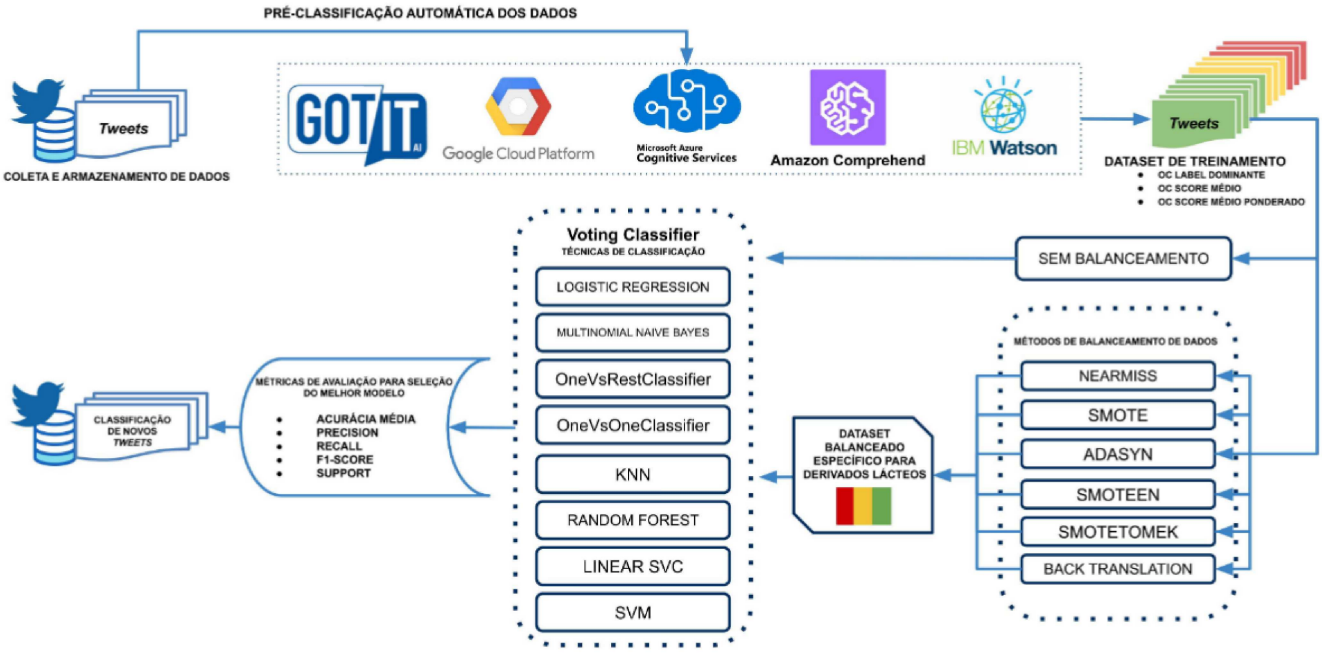
As manipulações e análises de dados foram realizadas no Python 3.11 (Van Rossum; Drake, 2009), com os pacotes Pandas 2.2.3 (McKinney, 2010) para manuseio dos dados, Matplotlib 3.10.1 (Hunter, 2007) para gerar gráficos, NLTK 3.9.1 (Bird; Loper, 2004) para remoção de *stopwords*, Scipy 1.15.2 (Virtanen *et al.*, 2020) e statsmodels 0.14.4 (Perktold; Seabold, 2024) para testes estatísticos, Wordcloud 1.9.4 (Mueller, 2012) para a nuvem de palavras, PyMySql 1.1.1 (Naoki, 2024) para utilização de SQL, além do Selenium 4.30 (Stewart *et al.*, s.d.) no processo de raspagem de dados nos sites.

Para a identificação de sexo e idade nas imagens de perfis, foram utilizados os modelos de visão computacional RetinaFace, opencv, yunet, mtcnn, e yolov8, inclusos na biblioteca Deepface 0.0.93 (Serengil; Ozpinar, 2020). Esses modelos utilizam técnicas de aprendizado de máquina multitarefas supervisionado e auto-supervisionado para a extração das características faciais (Yang *et al.*, 2024). O critério para definição do sexo foi a soma das porcentagens estimadas pelos modelos mencionados, e para a idade foi feita a média entre os valores estimados por esses mesmos modelos.

Para a análise de sentimentos das avaliações coletadas na raspagem de dados, foi utilizado o modelo desenvolvido por (Nogueira; Siqueira; Capriles, 2024). O modelo foi criado com *tweets* da rede social X contendo determinadas palavras-chave para categorias de produtos lácteos (bebidas lácteas, creme de leite, doce de leite, iogurte, leite, leite condensado, leite fermentado, manteiga, queijos, sorvetes), classificando-os com cinco APIs de PLN e categorizando-os de acordo com suas polaridades (positivas, neutras, negativas),

aplicando técnicas de balanceamento de dados quando necessário. Os dados então são classificados por diferentes técnicas e avaliados de acordo com métricas de desempenho, assim identificando o modelo de análise de sentimentos mais eficaz para a classificação de novos dados, com o processo metodológico apresentado na Figura 8.

Figura 8 – Processo metodológico da construção do modelo de análise de sentimentos.



Fonte: Nogueira, Siqueira e Capriles (2024).

Para comparativos entre categorias dos produtos de interesse obtidas no formulário e categorias de produtos coletados com a raspagem de dados, cada uma foi enquadrada em uma das 17 categorias a seguir: “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”.

5 RESULTADOS

5.1 FORMULÁRIO "ESTUDO DO COMPORTAMENTO DE BRASILEIROS SOBRE O CONSUMO DE PRODUTOS LÁCTEOS"

No formulário "Estudo do comportamento de brasileiros sobre o consumo de produtos lácteos", após a filtragem das respostas, o novo conjunto de dados passou a ter 838 registros, aproximadamente 94,47% do conjunto original (887 registros). O gênero feminino apresentou maior adesão de respondentes, com aproximadamente 63% do total, como mostra o Quadro 2. Os três participantes com gênero "Outros" foram removidos por representarem uma baixíssima parcela do total de participantes, levando o novo conjunto a ter 835 registros. Houve diferença estatística significativa entre os gêneros feminino e masculino (p-valor $\approx$ 0,00).

Quadro 2 – Distribuição do gênero dos respondentes do formulário.

Gênero	**Contagem**	**Proporção**
Feminino	526	62,77%
Masculino	309	36,87%
Outro	3	0,36%

Fonte: Elaborado pelo Autor (2025).

O Quadro 3 mostra a distribuição das faixas etárias dos respondentes dos formulários. A concentração está entre 25 e 44 anos e houve diferença significativa nas proporções entre todos os pares de faixa etária (p-valor $\approx$ 0,00 em todos os casos), exceto entre "18-24 anos" e "55-64 anos" (p-valor $\approx$ 0,81), e entre "25-34 anos" e "35-44 anos" (p-valor $\approx$ 0,66).

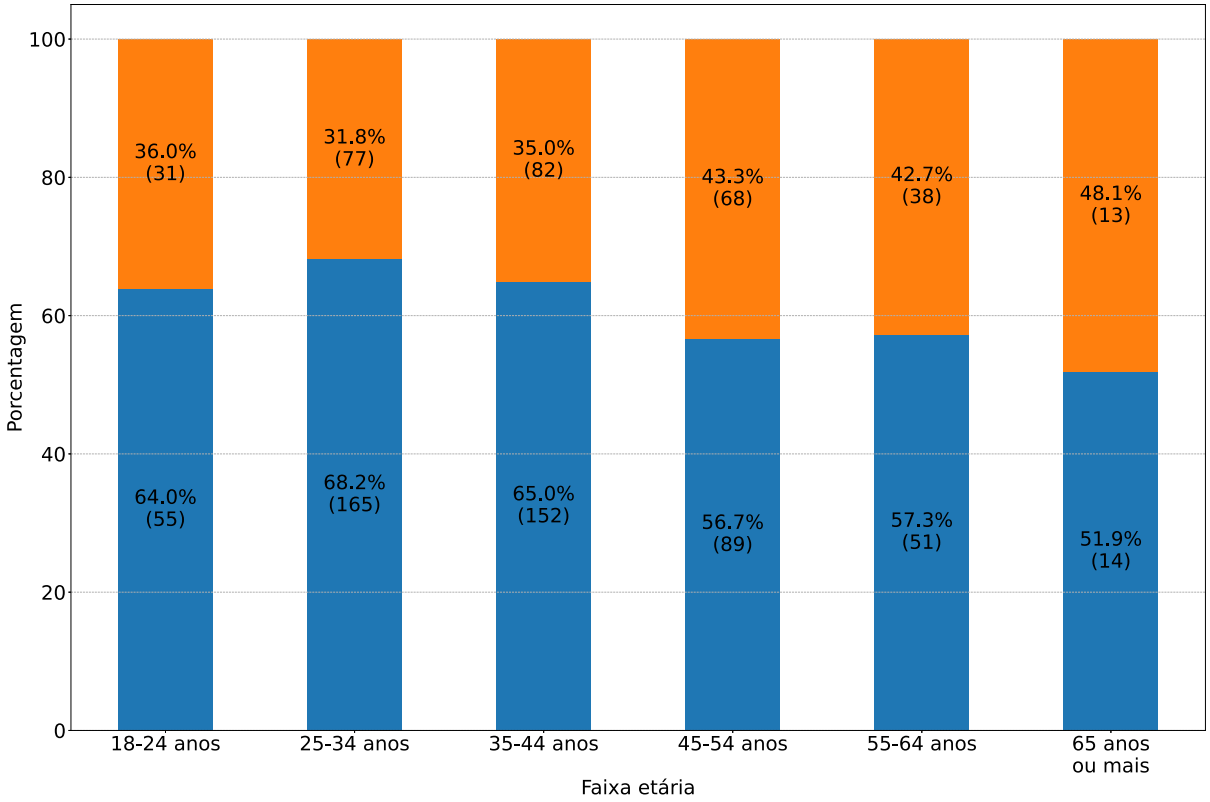
Quadro 3 – Distribuição das faixas etárias dos respondentes do formulário.

Faixa Etária	**Contagem**	**Proporção**
18-24 anos	86	10,30%
25-34 anos	242	28,98%
35-44 anos	234	28,02%
45-54 anos	157	18,80%
55-64 anos	89	10,66%
65 anos ou mais	27	3,22%

Fonte: Elaborado pelo Autor (2025).

Observando os gêneros por cada faixa etária, temos a distribuição da Figura 9, com participantes do gênero feminino sendo maioria em todas as faixas, principalmente entre 25 e 44 anos. As mesmas faixas etárias também representam grande parte dos participantes de gênero masculino. Não houve diferença estatística significativa entre as diferentes faixas etárias para as proporções de gênero (p-valor $\approx$ 0,12).

Figura 9 – Distribuição percentual de gênero por faixa etária dos participantes do formulário.



Valores percentuais e absolutos individuais em cada segmento por gênero: “Feminino” em azul, “Masculino” em laranja. Fonte: Elaborada pelo Autor (2025).

Para a renda familiar, a distribuição das respostas seguiu uma ordem crescente (Quadro 4). Pode ser que o formulário não tenha alcançado o público das rendas mais baixas tanto quanto as rendas mais altas, considerando que a renda familiar média do país foi estimada em R$2.846 para o ano de 2023 (IBGE, 2024). Não houve diferença estatística significativa apenas entre as respostas “Menos de R$ 1.000” e “Prefiro não responder” (p-valor $\approx$ 0,55), “R$ 1.000 - R$ 2.000” e “Prefiro não responder” (p-valor $\approx$ 0,10), “R$ 2.000 - R$ 5.000” e “R$ 5.000 - R$ 10.000” (p-valor $\approx$ 0,86).

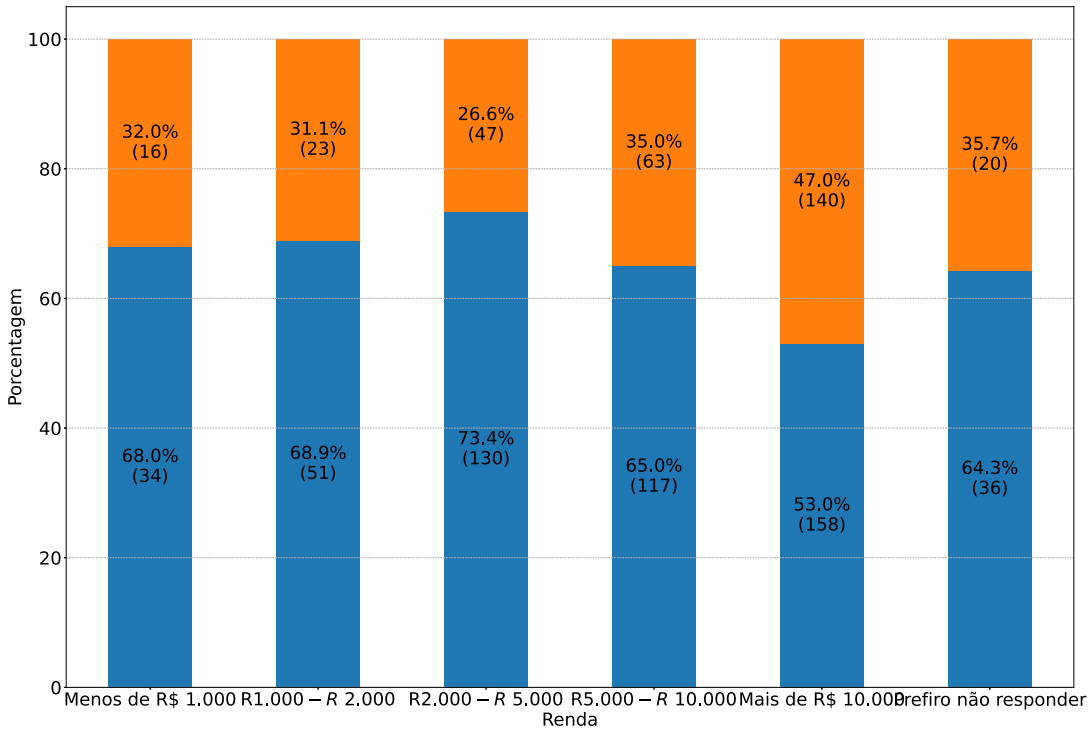
Quadro 4 – Distribuição das respostas para a faixa de renda dos respondentes.

Faixa de Renda	**Contagem**	**Proporção**
Menos de R$ 1.000	50	5,97%
R$ 1.000 - R$ 2.000	76	9,07%
R$ 2.000 - R$ 5.000	178	21,24%
R$ 5.000 - R$ 10.000	180	21,48%
Mais de R$ 10.000	298	35,56%
Prefiro não responder	56	6,68%

Fonte: Elaborado pelo Autor (2025).

A Figura 10 apresenta a distribuição de gênero dentro de cada faixa de renda. Participantes que se identificaram com o gênero feminino foram a maioria em todas as rendas, mas não houve diferença estatística significativa entre rendas e gêneros (p-valor ≈ 0,13).

Figura 10 – Distribuição de gênero por renda dos participantes do formulário.



Valores percentuais e absolutos individuais em cada segmento por gênero: “Feminino” em azul, “Masculino” em laranja. Fonte: Elaborada pelo Autor (2025).

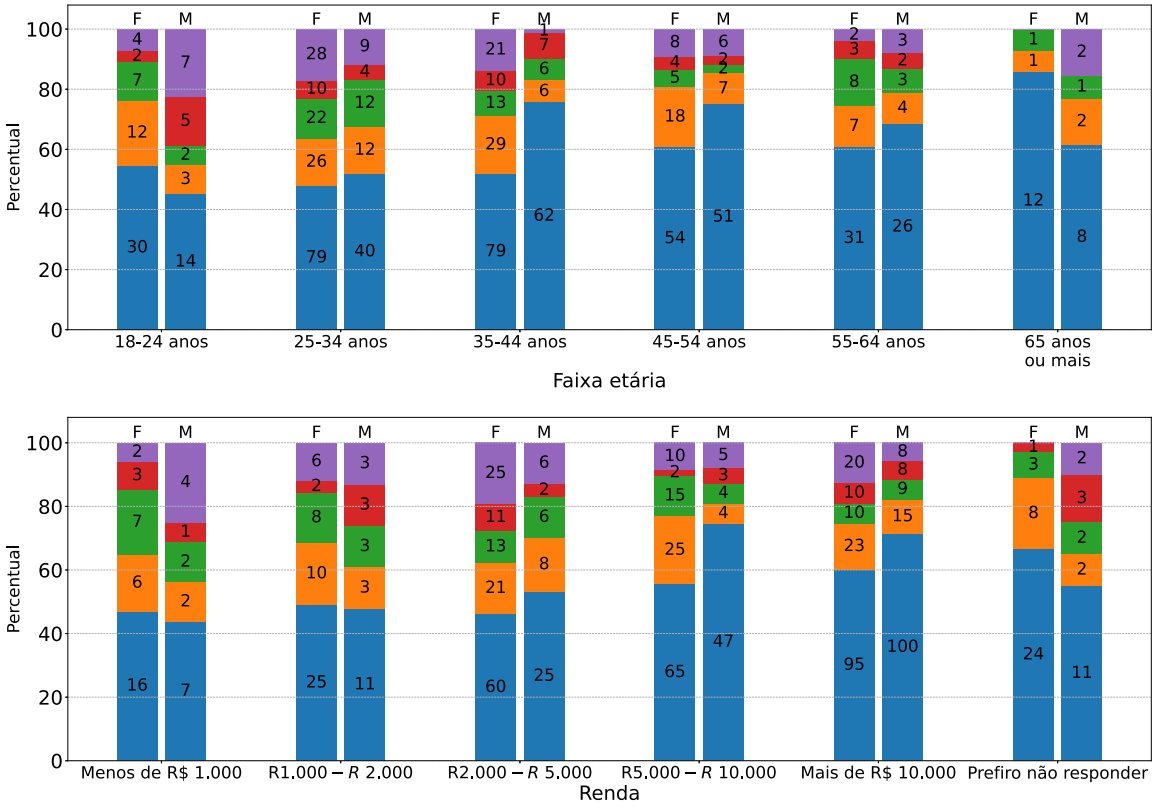
O termo “leitelho” não é comum para os participantes do formulário, ainda que ele seja produzido na mesma proporção da manteiga, essa sendo aproximadamente 80 mil toneladas por ano (Brasil, 2024). As respostas “Nada familiar” e “Pouco familiar” representam cerca de 73% das respostas para o nível de familiaridade com o termo (Quadro 5), com diferença significativa entre todas as respostas, exceto entre os níveis “Moderamente” e “Muito”.

Grau de Familiaridade	**Contagem**	**Proporção**
Nada familiar	488	58,23%
Pouco familiar	127	15,16%
Muito familiar	92	10,98%
Moderadamente familiar	82	9,79%
Bastante familiar	49	5,85%

Quadro 5 – Distribuição das respostas sobre a familiaridade com o termo “leitelho”.
Fonte: Elaborado pelo Autor (2025).

Para o nível de familiaridade com o termo por gênero em cada faixa etária (Figura 11), os níveis “Moderadamente familiar” e acima aparecem com maior frequência entre participantes de 18 a 34 anos. Nessas faixas, eles somam quase 30% das respostas. Por renda, os mesmos níveis chegam a quase 40% entre quem recebe até R$ 5.000,00. Entre mulheres com 65 anos ou mais, nenhuma respondeu “Bastante familiar” ou “Muito familiar”. Entre homens da mesma faixa etária, também não houve respostas para “Bastante familiar”. A análise estatística mostrou diferença significativa por gênero e faixa etária (p-valor $\approx$ 0,00). Homens de 18 a 24 anos foram mais frequentes em “Bastante” e “Muito familiar”. Já mulheres de 25 a 34 anos apareceram mais em “Muito familiar”. Nessa mesma faixa, o nível “Nada familiar” ficou abaixo do esperado pela sua proporção nas outras faixas etárias, enquanto entre homens foi o contrário. Também houve diferença por renda. Homens com renda acima de R$ 5.000 apareceram mais entre os que têm pouco conhecimento do termo. Para mulheres com renda entre R$ 2.000 e R$ 5.000, “Muito familiar” esteve acima do esperado, enquanto “Nada familiar” esteve abaixo em relação às demais rendas.

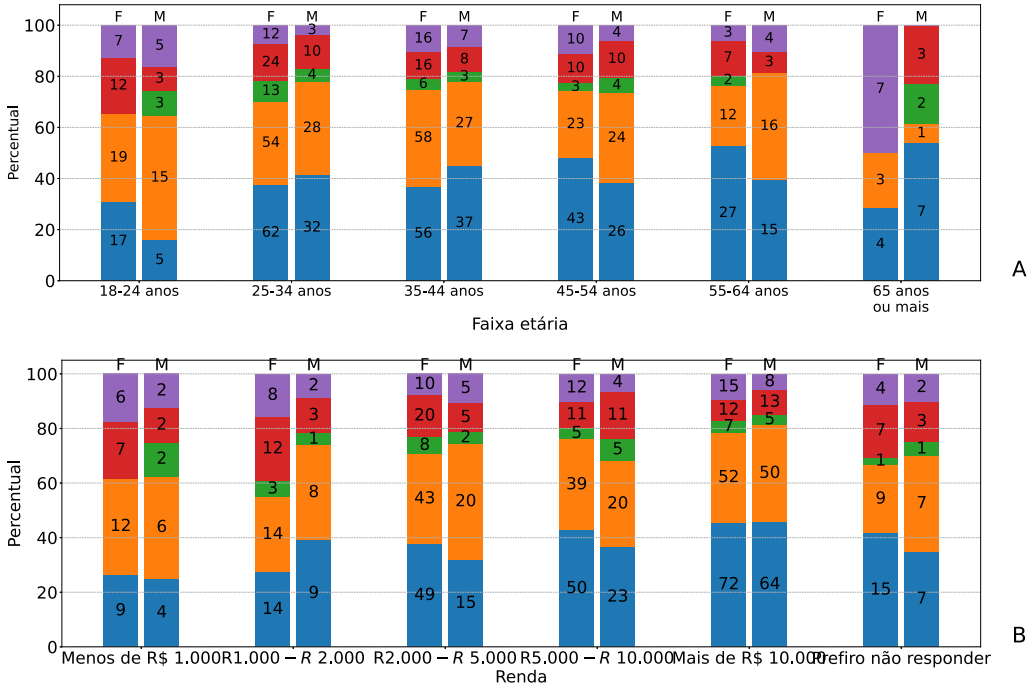
Figura 11 – Distribuição de familiaridade com o termo “leitelho” por gênero dos participantes do formulário, por (A) faixa etária e por (B) renda.



Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para a frequência de consumo dos produtos: “Nada familiar” em azul, “Pouco familiar” em laranja, “Moderadamente familiar” em verde, “Bastante familiar” em vermelho, “Muito familiar” em roxo. Fonte: Elaborada pelo Autor (2025).

Para a frequência de consumo desses produtos por faixa etária e por renda (Figura 12), existe o consumo diário em todos os grupos, apesar de o conhecimento do termo ter sido baixo, como mostrado anteriormente no Quadro 5. Exceto nos respondentes masculinos com 65 anos ou mais, houve diferença estatística significativa entre gêneros (p-valor $\approx$ 0,00) principalmente nessa faixa etária, com o consumo diário do gênero feminino estando muito acima das outras faixas, mesmo que esse mesmo grupo não tenha bastante nem muito conhecimento sobre o termo. Cerca de 40% dos respondentes de renda familiar de até R$ 2.000 consomem produtos à base de leitelho pelo menos uma vez na semana, ainda que nenhuma mulher com renda inferior a R$ 1.000 tenha escolhido essa resposta, mas não houve diferença significativa nos níveis de consumo por renda (p-valor $\approx$ 0,22), nem ao considerar os gêneros (p-valor $\approx$ 0,59). O consumo de laticínios por pessoa no Brasil entre 2017 e 2018 foi, na média, de 32,2 quilos, cerca de 0,09 quilos por dia, sendo consumidos principalmente leite, queijos e iogurte (Siqueira, 2021). Bebidas lácteas com leitelho, ainda que existam trabalhos que o usem como substituto do leite na produção, como em (Santos *et al.*, 2024), não são encontradas para compra de forma fácil em buscas na internet, apenas receitas caseiras, também podendo estar limitadas pelas questões de armazenamento e transporte.

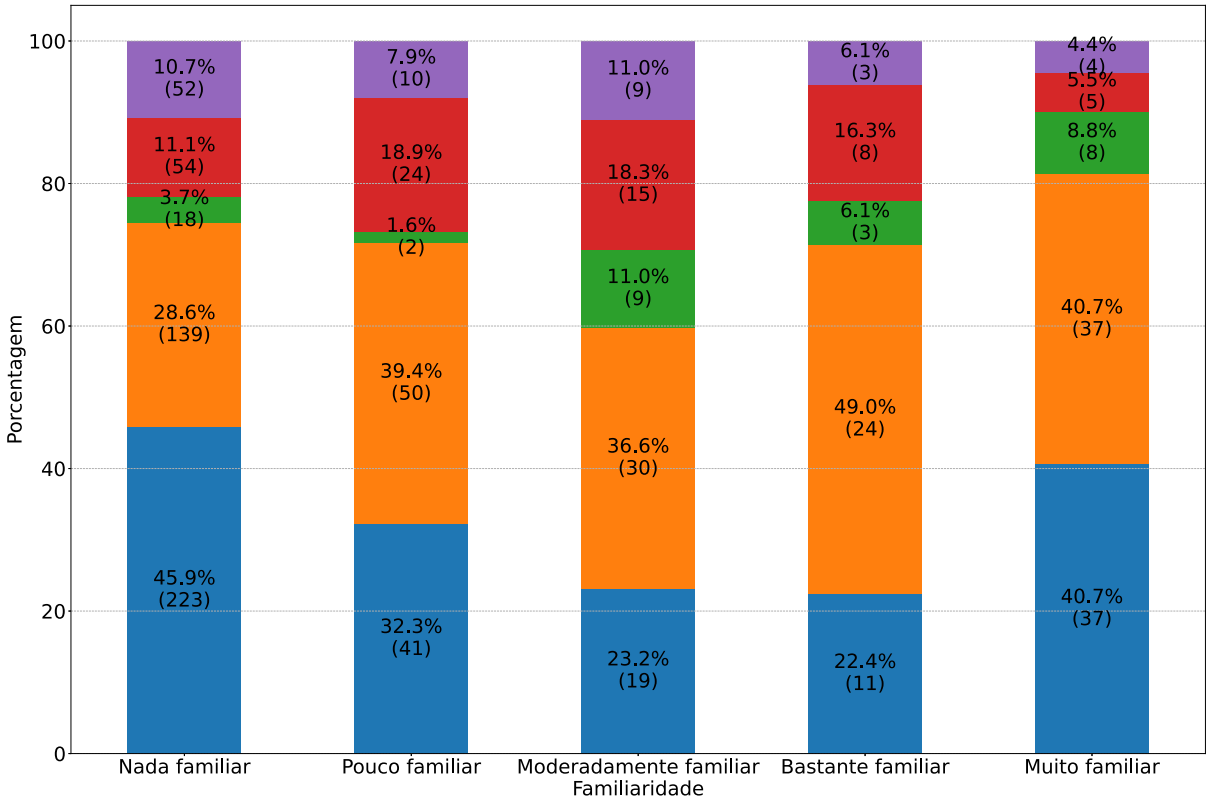
Figura 12 – Distribuição de consumo de produtos à base de leitelho por gênero dos participantes do formulário, por (A) faixa etária e por (B) renda



Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para a frequência de consumo dos produtos: “Nunca” em azul, “Raramente” em laranja, “Uma vez por semana” em verde, “Algumas vezes por semana” em vermelho, “Diariamente” em roxo. Fonte: Elaborada pelo Autor (2025).

Apesar da não familiaridade com o termo, existem participantes que responderam que consomem produtos à base de leitelho com diferentes frequências, até diariamente, mesmo tendo pouca ou nenhuma familiaridade com o termo, como mostra a Figura 13. Essa aparente incoerência pode ser explicada por diferentes fatores, como uma má interpretação da pergunta pelo participante, ou o participante considerar que algum produto que ele consome contém leitelho na composição, ou até uma produção caseira de manteiga em que o participante então utilize o leitelho resultante dessa produção.

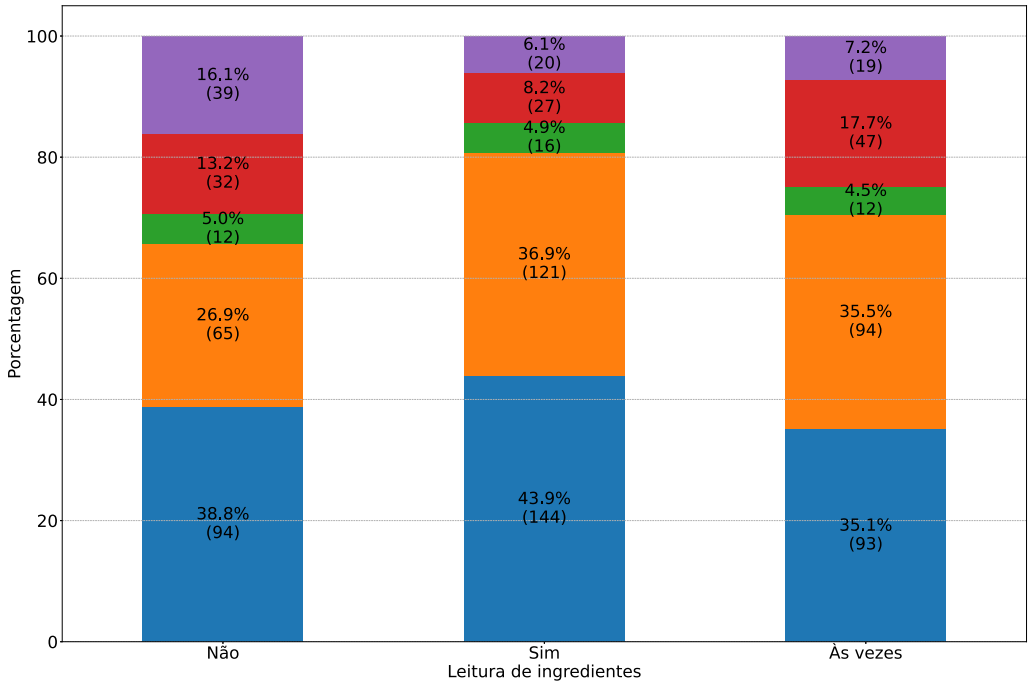
Figura 13 – Distribuição de consumo de produtos à base de leitelho por nível de familiaridade com o termo.



Para cada nível de familiaridade com o termo, os segmentos apresentam os valores percentuais e absolutos para a frequência de consumo dos produtos: “Nunca” em azul, “Raramente” em laranja, “Uma vez por semana” em verde, “Algumas vezes por semana” em vermelho, “Diariamente” em roxo. Fonte: Elaborada pelo Autor (2025).

Houve registros de participantes que consomem produtos lácteos mas não consultam a lista de ingredientes, ou o fazem apenas às vezes (Figura 14). Nesse caso, poderiam estar ingerindo leitelho sem perceber. A análise mostrou diferença estatística entre as respostas por frequência de consumo (p-valor $\approx 0{,}00$). Entre os que consomem diariamente, a resposta “Não” apareceu acima do esperado, enquanto a resposta “Sim” foi menor entre os que consomem algumas vezes ou todos os dias. Já a opção “Às vezes” foi bastante comum entre os que consomem produtos algumas vezes por semana.

Figura 14 – Distribuição de consumo de produtos à base de leitelho por leitura de lista de ingredientes de produtos lácteos.



Para cada resposta sobre leitura da lista de ingredientes, os segmentos apresentam os valores percentuais e absolutos para a frequência de consumo dos produtos: “Nunca” em azul, “Raramente” em laranja, “Uma vez por semana” em verde, “Algumas vezes por semana” em vermelho, “Diariamente” em roxo. Fonte: Elaborada pelo Autor (2025).

O Quadro 6 apresenta as dez respostas que mais apareceram nas quatro perguntas sobre “palavra, sentimento e/ou emoção” sobre leitelho. Das 3.353 respostas, 1.030 foram distintas, enquanto as dez respostas mais comuns correspondem a 21,86% do total. Das dez respostas mais frequentes, três foram adjetivos (“cremoso”, “nutritivo”, “saboroso”) sendo que nenhuma dessas expressa sentimento negativo. Também foi disponibilizada uma imagem do coproduto antes dessas perguntas no formulário.

Quadro 6 – Dez palavras mais citadas e suas respectivas frequências e porcentagens.

Palavra	**Frequência**	**Porcentagem**
leite	105	3,13%
cremoso	93	2,77%
iogurte	91	2,71%
curiosidade	83	2,48%
sabor	72	2,15%
branco	64	1,91%
nutritivo	62	1,85%
impressão	58	1,73%
saboroso	57	1,70%
gordura	50	1,49%

Fonte: Elaborada pelo Autor (2025).

Com as mesmas respostas, agora sem possíveis *stopwords* e separando por palavras, temos um total de 4.103 palavras, das quais 787 são distintas. A Figura 15 apresenta as palavras após essa remoção. A palavra "parece" foi muito presente, sendo usada em respostas como "parece bom", "parece coalhada", ou "parece requeijão".

Figura 15 – Nuvem de palavras usadas para descrever o leitelho.

Fonte: Elaborada pelo Autor (2025).

Foi detectado pelo menos um adjetivo em 1.844 respostas (54,78% do total), assim sendo possível classificar a polaridade delas. O Quadro 7 traz essas respostas em que foi possível analisar o sentimento delas, sendo 234 (12,68%) com polaridade negativa, 782 (42,41%) neutra, e 828 (44,90%) positiva, havendo diferença significativa entre todas as proporções (p-valor $\approx$ 0,00). Mesmo que a palavra "ruim" tenha aparecido em diferentes frases na polaridade negativa (como "gosto ruim" ou "sabor ruim"), essas respostas somam pouco mais de 5% da polaridade negativa, que por sua vez representa menos de 13% do total das polaridades das respostas. Também é possível que o adjetivo "Azedo" não seja no sentido negativo, como para se referir ao gosto ácido do limão, podendo reduzir ainda mais o total de sentimentos negativos sobre o leitelho.

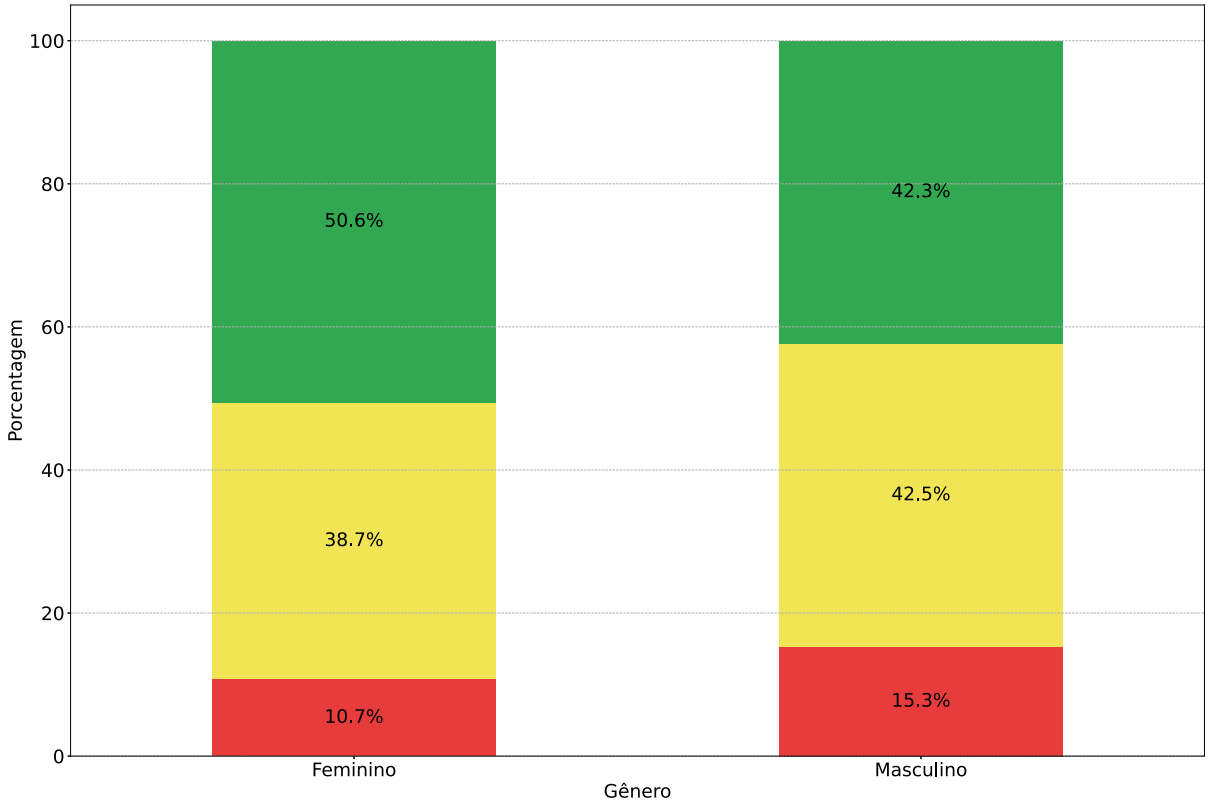
Ao analisar as polaridades por gênero (Figura 16), a positiva foi maioria apenas para o grupo feminino, enquanto no masculino foi a neutra, podendo indicar que as mulheres têm uma percepção mais positiva que os homens sobre esse coproduto. Ainda que a polaridade neutra tenha sido maioria para o gênero masculino, a positiva também representa mais de 40% no total analisado para esse gênero, com diferença estatística significativa (p-valor = 0,001) entre os gêneros nas polaridades positiva e negativa.

Quadro 7 – Classificação de palavras por polaridade e suas respectivas porcentagens.

Polaridade Negativa	**Polaridade Neutra**	**Polaridade Positiva**
Azedo* (17,09%)	Branco (07,94%)	Cremoso (11,22%)
Gorduroso* (10,68%)	Coalhada (04,09%)	Nutritivo (07,48%)
Estranho* (09,40%)	Pastoso (03,85%)	Saboroso (06,88%)
Ruim* (06,83%)	Diferente (03,10%)	Gostoso (06,03%)
Gosto ruim (03,42%)	Normal (02,98%)	Bom* (04,46%)
Insosso (03,42%)	Indiferente (02,36%)	Interessante (04,46%)
Inferior (01,71%)	Desconhecido (02,10%)	Saudável (03,86%)
Aguado (01,71%)	Leve (01,99%)	Agradável (03,38%)
Sabor ruim (01,71%)	Doce (01,86%)	Consistente (02,90%)
Fraco (01,71%)	Curioso (01,74%)	Bonito (02,65%)

As palavras acompanhadas por "*" também estiveram entre as mais frequentes nas avaliações coletadas na raspagem de dados, apresentadas posteriormente no Quadro 9. Fonte: Elaborado pelo Autor (2025).

Figura 16 – Distribuição percentual das polaridades das respostas sobre leitelho por gênero.

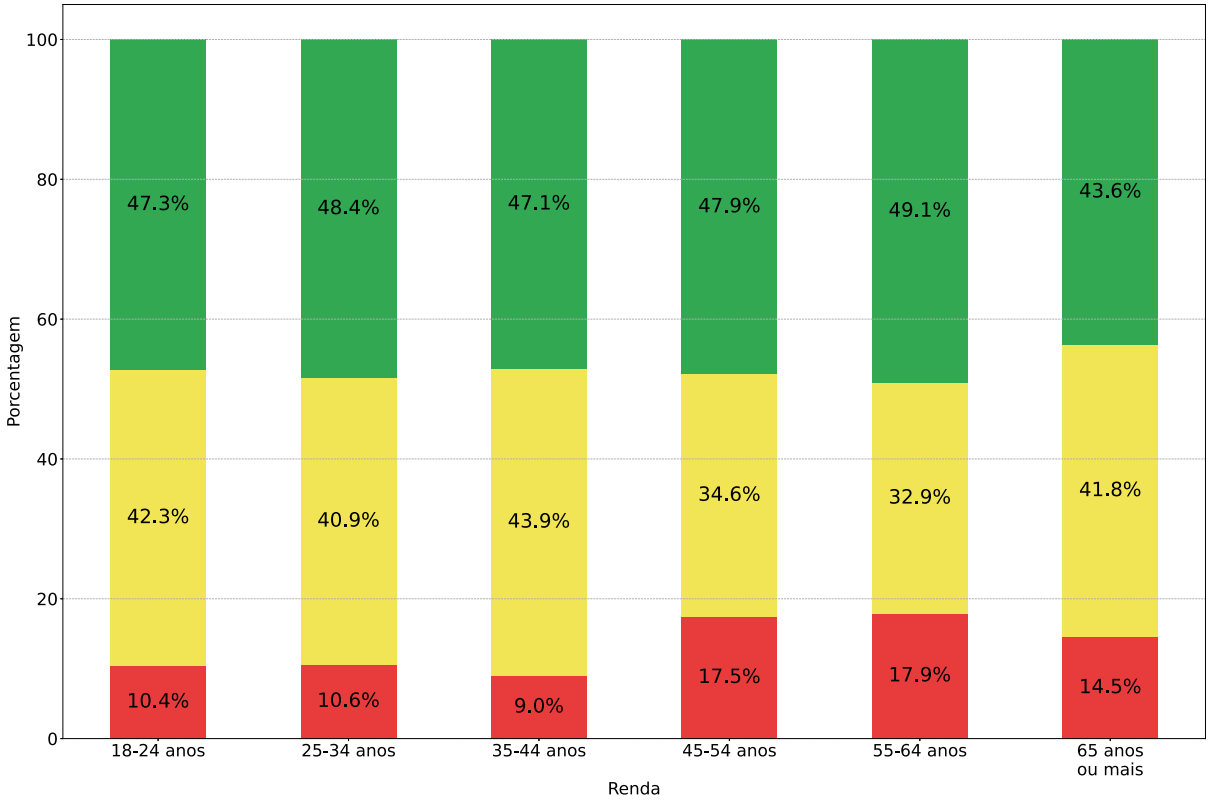


Para cada gênero, os valores percentuais das polaridades são apresentados, sendo verde para positivas, amarelo para neutras, vermelho para negativas. Fonte: Elaborada pelo Autor (2025).

Ao analisar as polaridades por faixa etária (Figura 17), é possível ver a diferença significativa de polaridade negativa nas faixas etárias de 45 a 64 anos (p-valor $\approx$ 0,00 nas

duas faixas). Considerando a polaridade negativa mais alta que nas demais faixas, pode ser de interesse que haja uma campanha de informação sobre o leitelho e seus produtos, junto com seus benefícios, mais adequado para pessoas com 45 anos ou mais. Ainda que a polaridade positiva não tenha superado 50% do total em nenhum grupo, ela foi a maioria em todos eles.

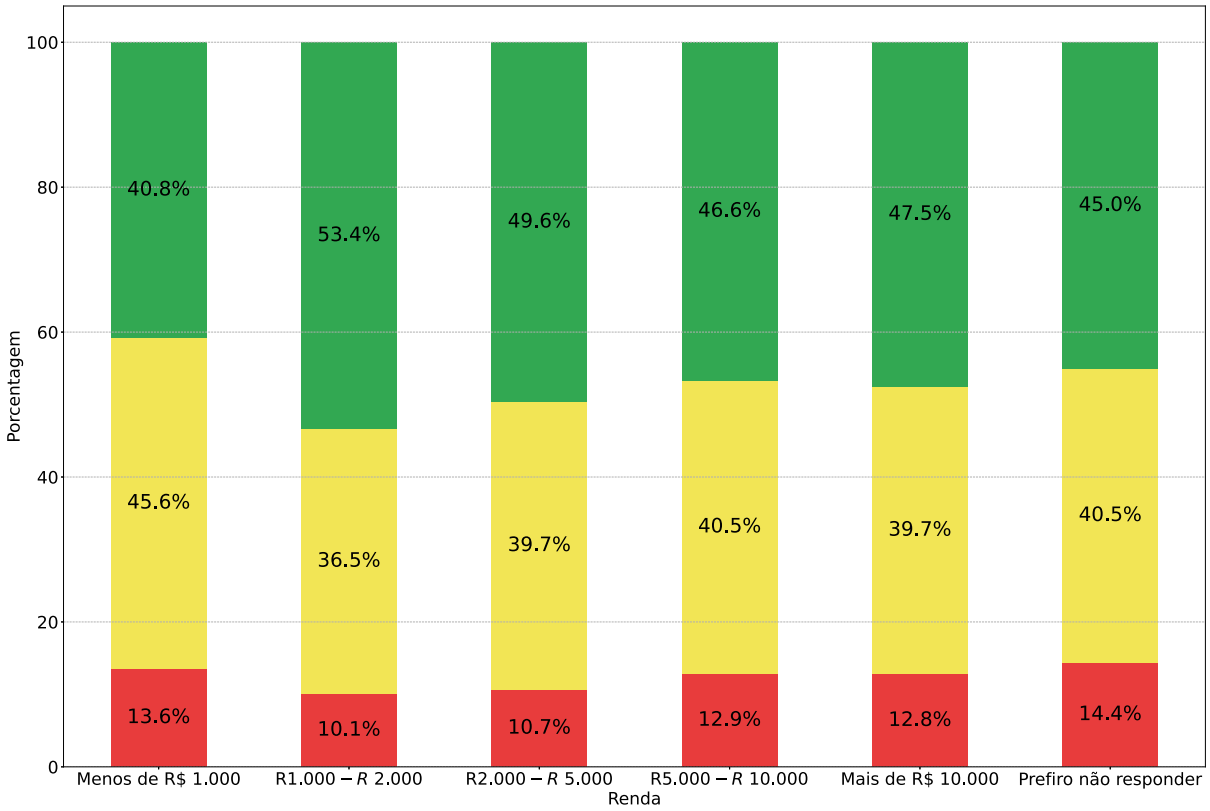
Figura 17 – Distribuição percentual das polaridades das respostas sobre leitelho por faixa etária.



Para cada faixa etária, os valores percentuais das polaridades são apresentados, sendo verde para positivas, amarelo para neutras, vermelho para negativas. Fonte: Elaborada pelo Autor (2025).

Já ao analisar por renda familiar (Figura 18), não houve diferença estatística significativa entre os grupos (p-valor $\approx$ 0,40), nem ao separar por gênero (p-valor $\approx$ 0,07), ainda que o grupo com renda inferior a R$1.000 tenha apresentado mais polaridades neutras e negativas que os demais. A menor porcentagem de polaridade positiva também é nesse grupo, enquanto o grupo de renda de R$1.000 a R$2.000 respondeu de forma positiva em mais de 50% do total, sendo a maior porcentagem entre gênero, faixa etária, ou renda familiar.

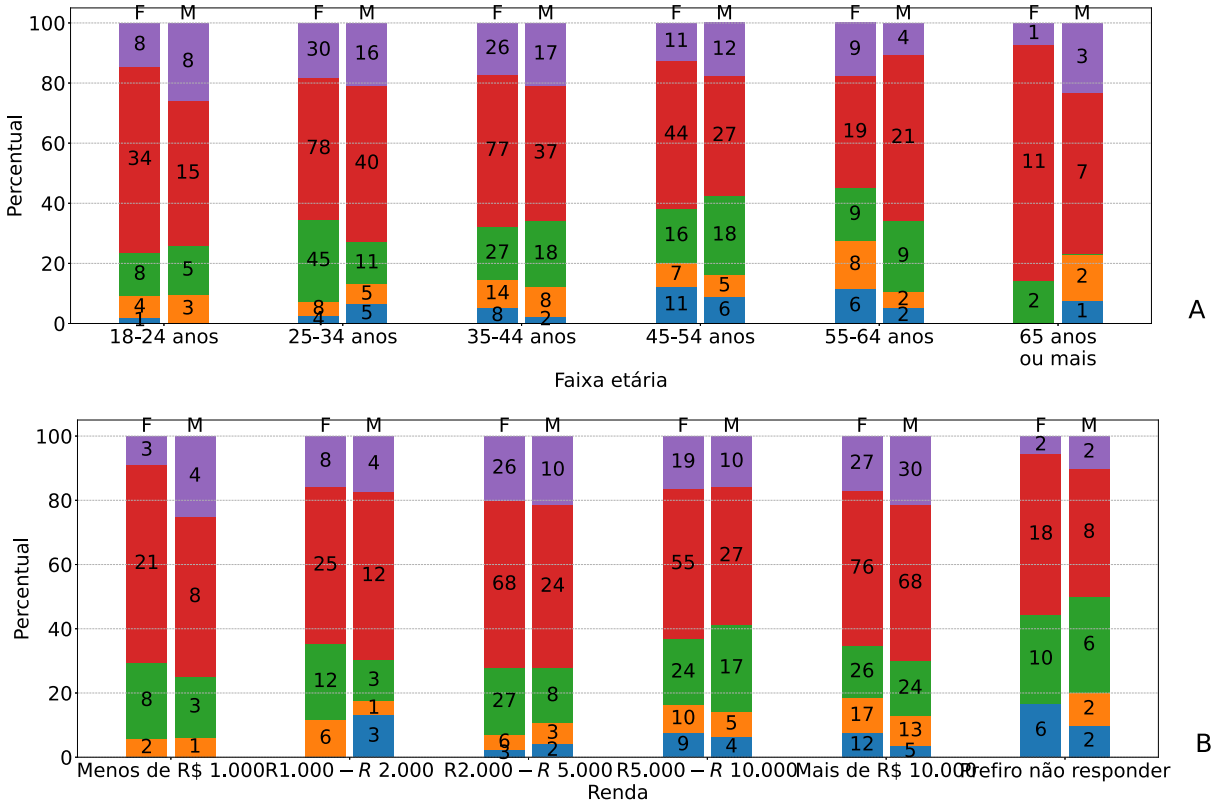
Figura 18 – Distribuição percentual das polaridades das respostas sobre leitelho por renda.



Para cada renda, os valores percentuais das polaridades são apresentados, sendo verde para positivas, amarelo para neutras, vermelho para negativas. Fonte: Elaborada pelo Autor (2025).

Quando perguntados sobre o interesse em experimentar produtos à base de leitelho, as respostas positivas (“Provavelmente sim” e “Definitivamente sim”) são maioria em todas as faixas etárias e também em todas as rendas, como mostra a Figura 19, e, ao considerar também as polaridades majoritariamente positivas nas respostas para “palavra, sentimento e/ou emoção”, mostrando como produtos com leitelho podem ser bem recebidos se bem apresentados. Apesar da resposta “Provavelmente sim” ser relativamente mais presente para respondentes maiores de 65 anos do que nas demais faixas, esse grupo representa pouco mais de 3% dos participantes do formulário, não havendo diferença significativa entre faixas etárias (p-valor $\approx$ 0,09), nem ao considerar os gêneros (p-valor $\approx$ 0,10). Na distribuição por renda, não houve nenhuma resposta “Definitivamente não” entre participantes com renda menor que R$1.000. As respostas “Definitivamente sim” e “Neutro” alternam como a 2ª resposta mais escolhida, exceto na renda “Prefiro não responder”, onde o “Definitivamente não” teve o dobro de respostas de “Definitivamente sim”. Também não houve diferença significativa nessas respostas entre os grupos de renda (p-valor $\approx$ 0,06), nem ao considerar os gêneros (p-valor $\approx$ 0,31).

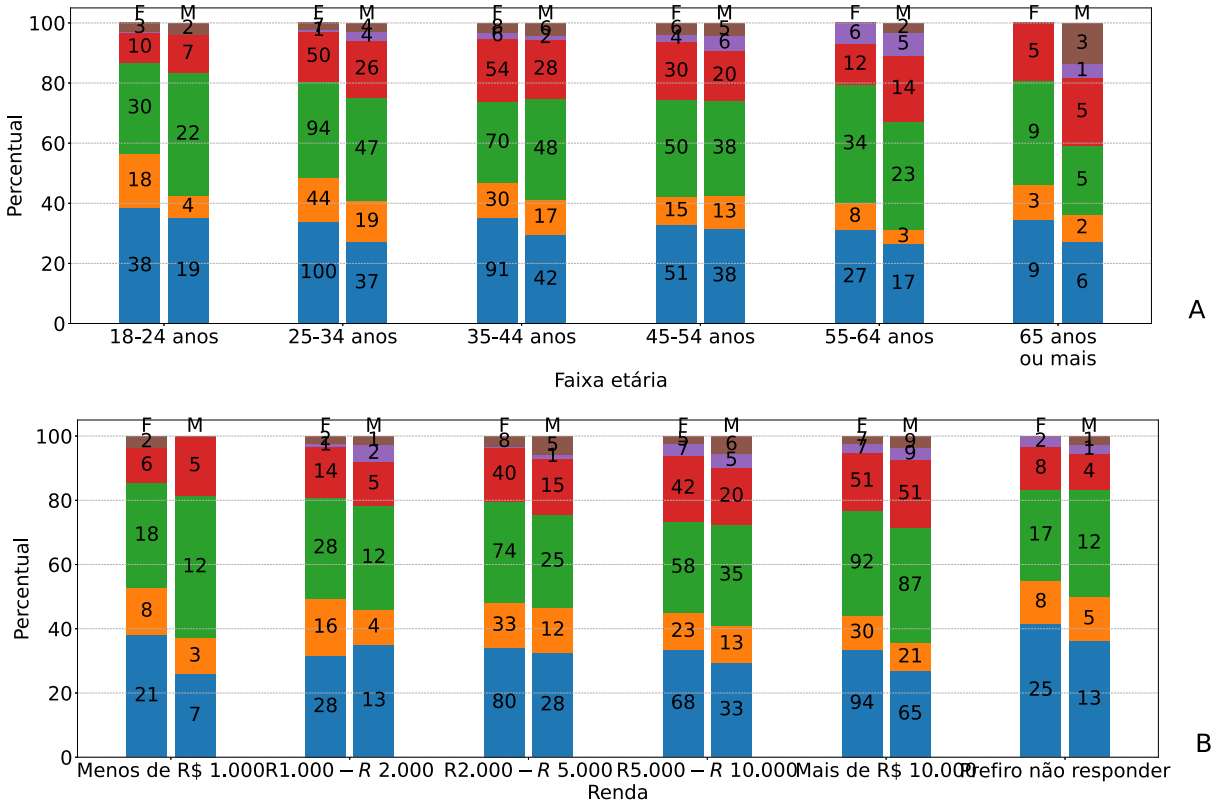
Figura 19 – Distribuição percentual da vontade de experimentar alimentos contendo leitelho por (A) faixa etária e por (B) renda.



Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para a vontade de experimentar esses alimentos: “Definitivamente não” em azul, “Provavelmente não” em laranja, “Neutro” em verde, “Provavelmente sim” em vermelho, “Definitivamente sim” em roxo. Fonte: Elaborada pelo Autor (2025).

E para os produtos que essas pessoas teriam mais interesse em experimentar, os principais são “produtos lácteos fermentados”, “pães, bolos e biscoitos” (panificação), “Chocolates”, “Produtos lácteos” (fermentados ou não), “Produtos cárneos”, entre outros, também seguindo a mesma tendência nas diferentes faixas etárias (Figura 20), não havendo diferença estatística significativa nos produtos entre faixa etária por gênero nem por renda. Ao desconsiderar o gênero, houve menos adesão na opção “Bebidas lácteas não fermentadas com alto teor de proteínas” nos participantes de até 24 anos, e mais interesse por “Chocolates” pelos participantes entre 25 e 34 anos. Existem trabalhos que mostram os benefícios para as propriedades reológicas e nutricionais com a utilização do leitelho na fabricação de pães, como (Al-Jahani, 2017), novamente mostrando o benefício do leitelho na composição, mas os resultados de buscas na *internet* não trazem produtos nacionais com esse ingrediente, apenas receitas caseiras.

Figura 20 – Distribuição de produtos de interesse por gênero dos participantes do formulário, por (A) faixa etária e por (B) renda.

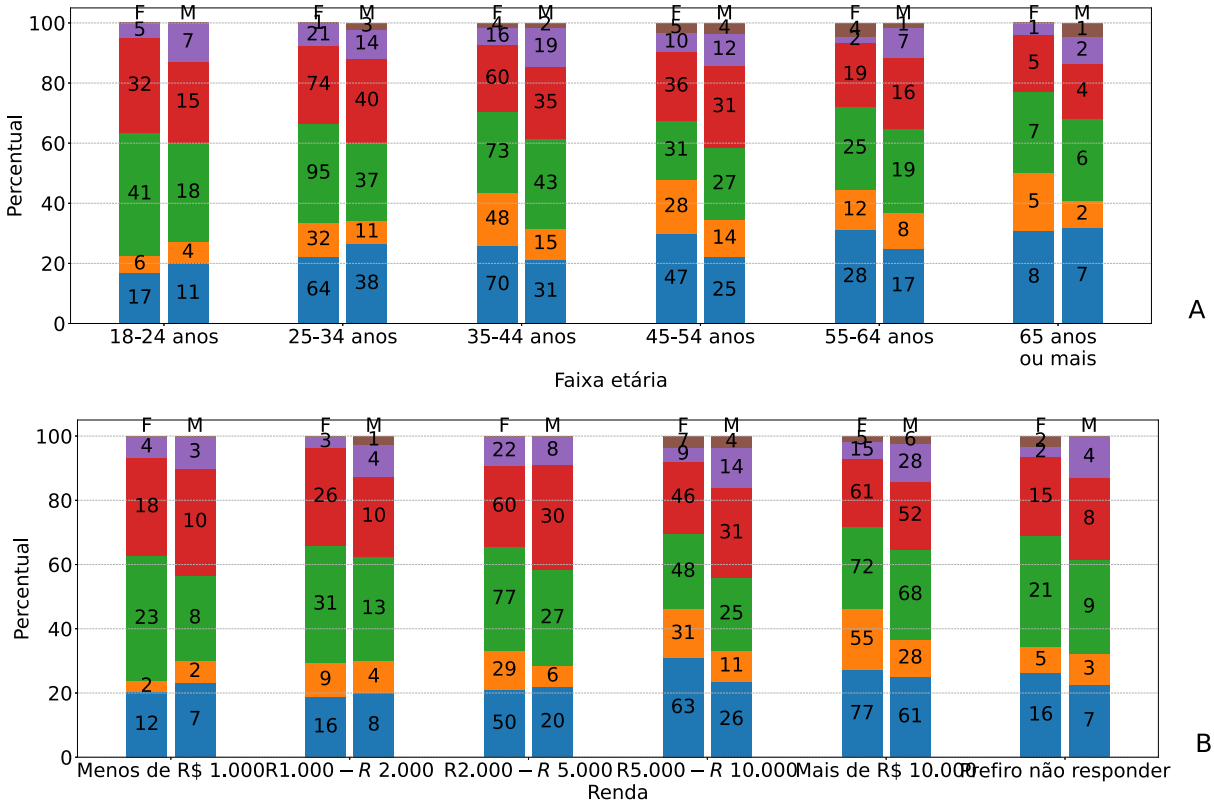


Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para os produtos de interesse: “Pães, bolos e biscoitos” em azul, “Chocolates” em laranja, “Produtos lácteos fermentados (como iogurte ou bebidas lácteas)” em verde, “Bebidas lácteas não fermentadas com alto teor de proteínas” em vermelho, “Produtos cárneos (como mortadela ou salame)” em marrom, “Outros” em roxo. Fonte: Elaborada pelo Autor (2025).

Os principais motivos que levariam as pessoas a experimentar produtos contendo leitelho também seguem as mesmas proporções nas diferentes idades e rendas (Figura 21), com as principais respostas sendo curiosidade, sabor e valor nutricional. Esses motivos foram mais presentes do que sustentabilidade ambiental e preço, com diferença estatística significativa entre as faixas por gênero (p-valor $\approx 0{,}02$). No sexo feminino de até 34 anos, foi mais aderente à opção “Curiosidade”, enquanto na faixa seguinte predomina a opção “Sustentabilidade ambiental” no mesmo gênero.

O preço baixo não foi predominante em nenhuma faixa etária, e ainda foi o menos significativo para o gênero feminino entre 55 e 64 anos. Já ao avaliar a diferença por renda (p-valor $\approx 0{,}00$), o preço baixo é significativo para o gênero masculino nas rendas acima de R$ 5.000. O sabor como motivador não foi significativamente diferente entre nenhum grupo de faixa etária ou renda. Esses motivos podem ser norteadores para os produtores entenderem o que os clientes buscarão para comprar os produtos contendo o leitelho.

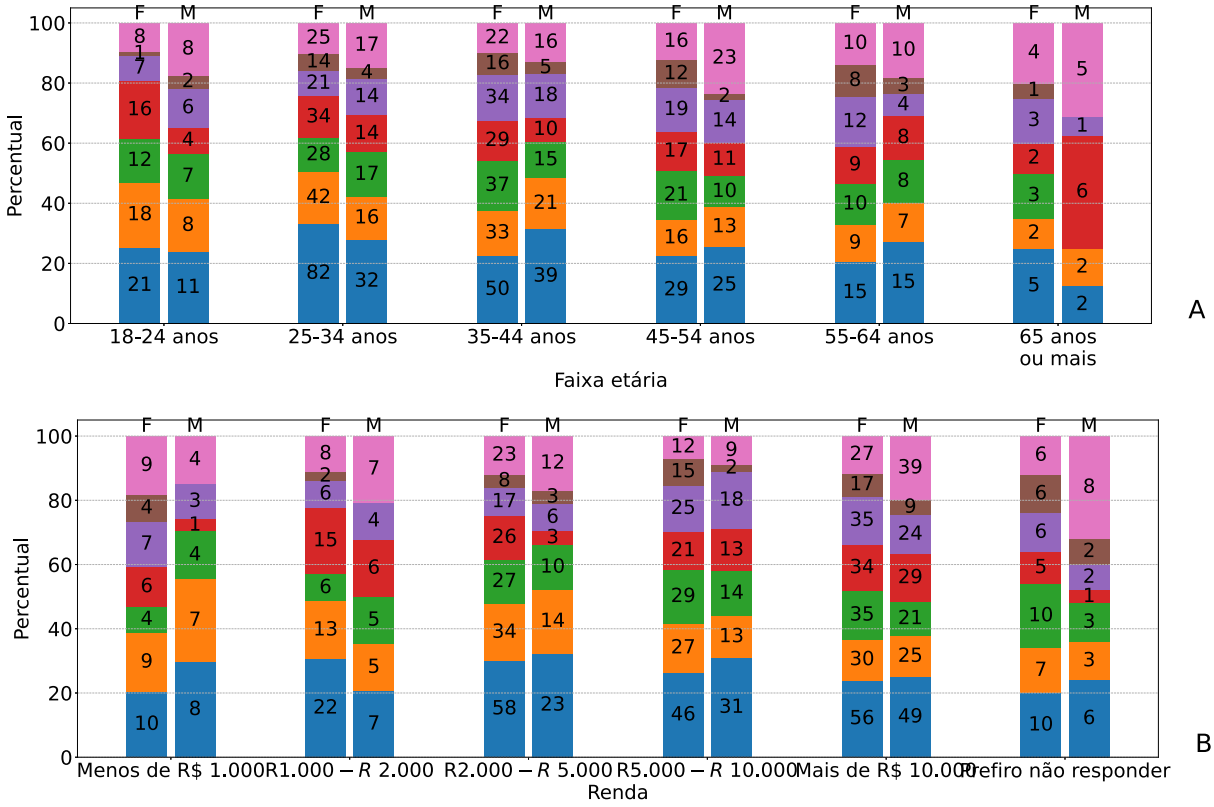
Figura 21 – Distribuição dos motivos para experimentar produtos à base de leitelho por gênero dos participantes do formulário, por (A) faixa etária e por (B) renda.



Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para os motivos para experimentar esses produtos: “Valor nutricional” em azul, “Sustentabilidade ambiental” em laranja, “Curiosidade” em verde, “Sabor” em vermelho, “Preço baixo” em roxo, “Outros” em marrom. Fonte: Elaborada pelo Autor (2025).

Nos motivos para não experimentar esses produtos (Figura 22), as respostas mais frequentes foram “sabor ruim”, “textura ruim”. Não houve respostas agrupadas em “Outros” para o gênero masculino com renda inferior a R$ 2.000. Entre as respostas dos motivos para não experimentar, também existia a opção “Nenhum”, sendo a 4ª opção mais escolhida, com pouco mais de 13% de presença nas respostas. Apesar das respostas “Presença de substâncias nocivas à saúde humana” e “Qualidade nutricional inferior” estarem presentes entre as 4ª ou 5ª mais escolhidas, essas não são informações que refletem a realidade, dado que o leitelho tem valores nutricionais próximos ao leite desnatado, além de mais fosfolipídeos, que possuem propriedades que os tornam bons agentes emulsificantes e potenciais anticarcinogênicos, e também alguns ácidos graxos que podem auxiliar na melhora de disfunções cognitivas causadas pelo Mal de Parkinson (Santa Rosa; Pires, 2021). Também não houve diferença estatística significativa por faixa etária (p-valor $\approx$ 0,13) nem por renda (p-valor $\approx$ 0,09) entre gêneros para esses motivos.

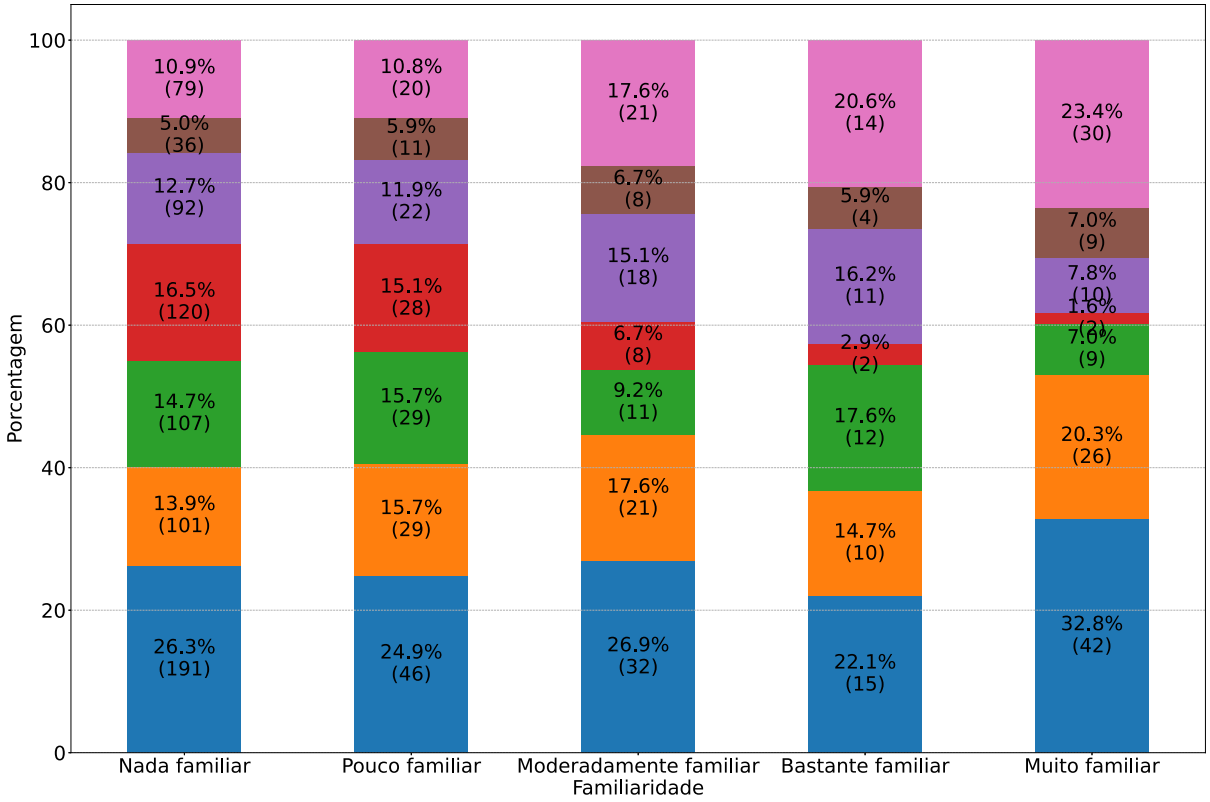
Figura 22 – Distribuição de motivos para não experimentar produtos à base de leitelho por gênero dos participantes do formulário, por (A) faixa etária e por (B) renda.



Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para os motivos para não experimentar esses produtos: “Sabor ruim” em azul, “Textura ruim” em laranja, “Presença de substâncias não adequadas à saúde humana” em verde, “Nenhum ou pouco conhecimento sobre leitelho” em vermelho, “Qualidade nutricional inferior” em roxo, “Nenhum” em rosa, “Outros” em marrom. Fonte: Elaborada pelo Autor (2025).

Curiosamente, “sabor ruim” é a resposta mais ocorrente até para quem tem pouca ou nenhuma familiaridade com o termo “leitelho” (Figura 23), indicando um pré-conceito negativo com o produto, somando cerca de 40% das respostas, mostrando novamente a importância da informação sobre o leitelho e suas propriedades.

Figura 23 – Distribuição dos motivos para não consumir produtos à base de leitelho por nível de familiaridade com o termo.



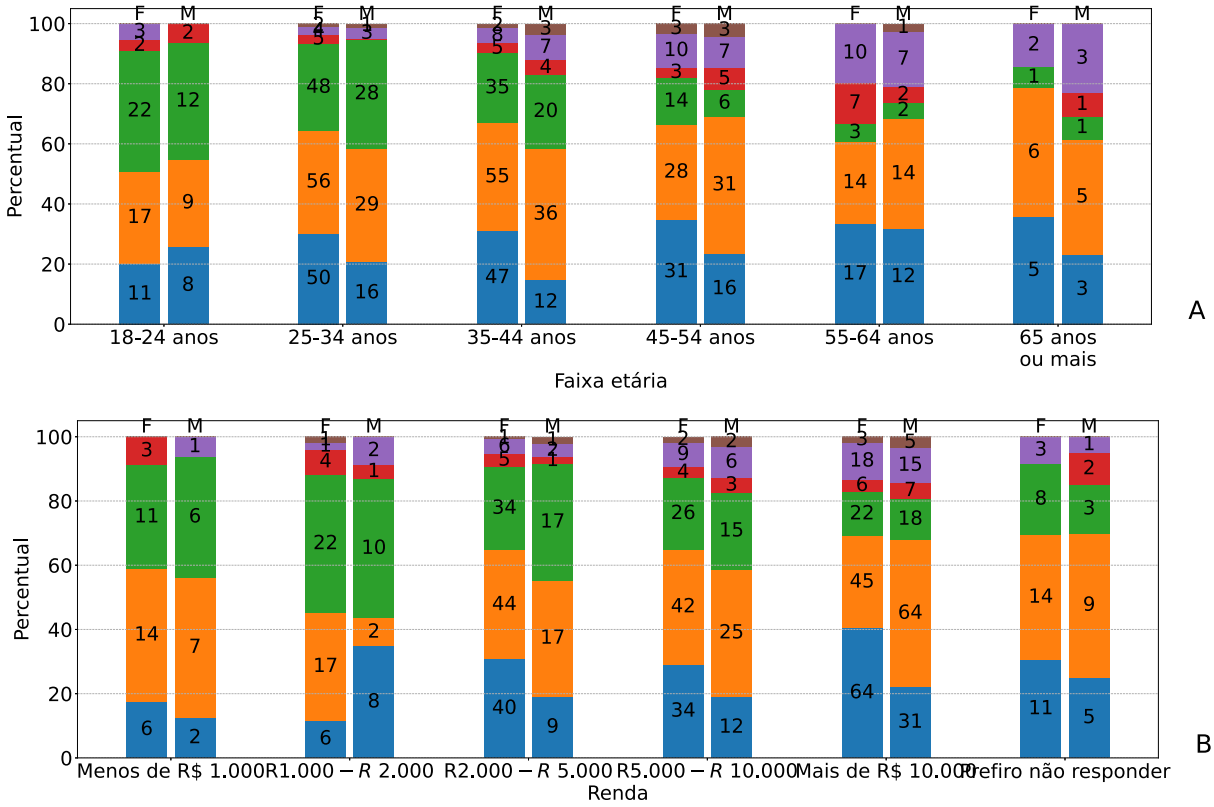
Para cada nível de familiaridade, os segmentos apresentam os valores percentuais e absolutos para os motivos para não experimentar esses produtos: “Sabor ruim” em azul, “Textura ruim” em laranja, “Presença de substâncias não adequadas à saúde humana” em verde, “Nenhum ou pouco conhecimento sobre leitelho” em vermelho, “Qualidade nutricional inferior” em roxo, “Nenhum” em rosa, “Outros” em marrom. Fonte: Elaborada pelo Autor (2025).

As fontes de informação acessadas pelos participantes sobre alimentos se concentram, no geral, na *internet* e profissionais da área, sendo que as redes sociais estão mais presentes entre os respondentes com menos de 35 anos, diminuindo a ocorrência nas idades superiores (Figura 24). Para as faixas etárias, houve diferença estatística significativa (p-valor $\approx$ 0,00), com “Redes sociais (Instagram e Facebook)” predominante entre os participantes de até 34 anos, enquanto foi o menos significativo a partir dos 45 anos.

Também houve diferença significativa entre os gêneros de renda (p-valor $\approx$ 0,00), novamente com as redes sociais mais associadas aos participantes com renda até R$ 5.000, e “Profissionais da área” mais associados ao gênero feminino com renda acima de R$ 10.000. Não houve nenhuma resposta “Profissionais sem formação especializada na área” para as faixas etárias “18–24 anos” e “Maior de 65 anos”, nem para as rendas “Menos de R$ 1.000” e “Prefiro não responder”, essa sendo a resposta menos escolhida nos grupos em que apareceu.

A televisão também não foi muito além de 10% das respostas, mas com forte associação ao gênero feminino entre 55 e 64 anos. Os meios de informação já são usados para *marketing* em geral, então podem ser usados para a venda de produtos à base de leitelho e também para informar corretamente sobre os benefícios desse coproduto, visando atrair esses consumidores.

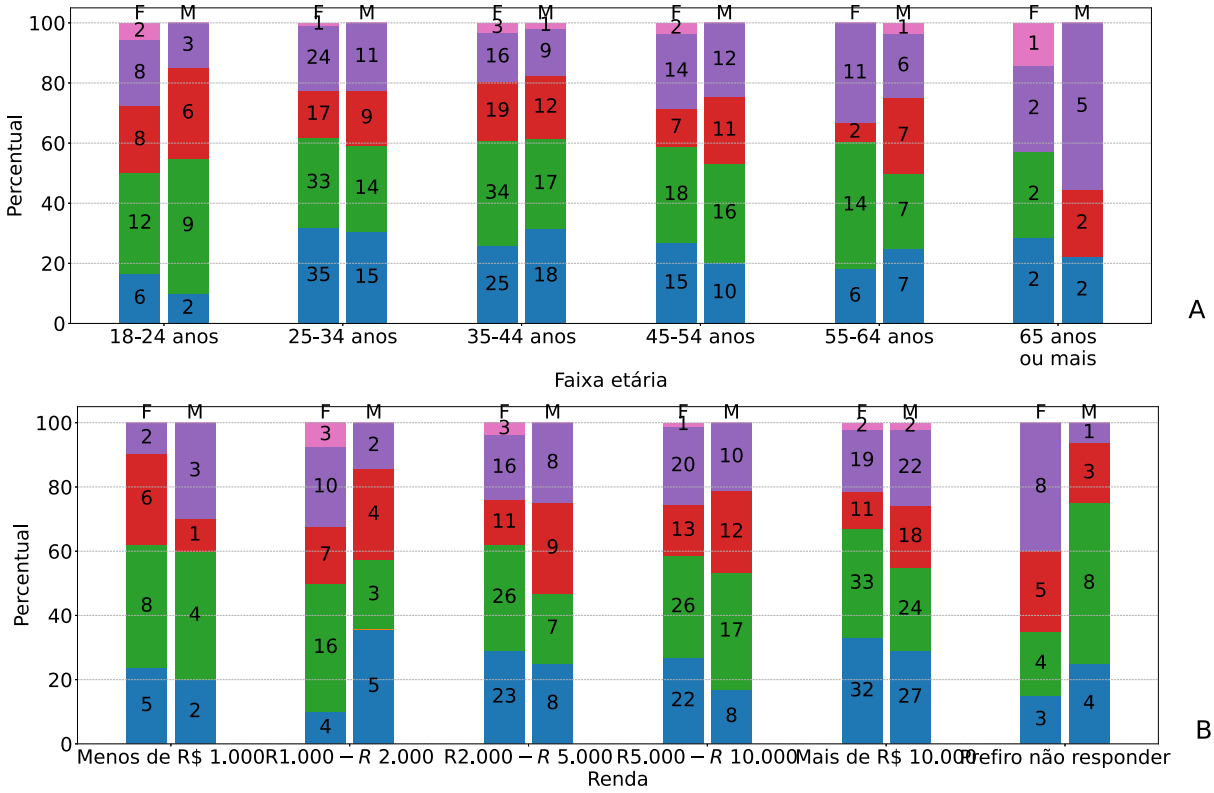
Figura 24 – Distribuição dos meios de informação por gênero dos participantes do formulário, por (A) faixa etária e por (B) renda.



Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para os meios de informação: “Profissionais da área” em azul, “Sites da internet” em laranja, “Redes sociais (Instagram ou Facebook)” em verde, “Televisão” em vermelho, “Profissionais sem formação especializada na área” em marrom, “Outros” em roxo. Fonte: Elaborada pelo Autor (2025).

Quando perguntados se não confiam em novos alimentos, as respostas também foram proporcionais entre os grupos (Figura **??**), não havendo diferença significativa entre os gêneros por faixa etária (p-valor $\approx 0{,}42$) nem por renda (p-valor $\approx 0{,}31$). Em todas as faixas de renda (exceto entre R$ 1.000 e R$ 2.000), a maioria discorda, em qualquer nível, de não confiar, mesmo que não seja significativo. Discordar totalmente da afirmação foi a resposta que menos apareceu em todos os casos, também sem ocorrências nas rendas “Menos de R$ 1.000” e “Prefiro não responder”. Considerando todos os três níveis de discordância com a afirmação, essas três respostas representam de 50 a 60% do total.

Figura 25 – Distribuição da não confiança em novos produtos por gênero dos participantes do formulário, por (A) faixa etária e por (B) renda.



Para cada faixa etária e cada renda, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para o nível de concordar com a afirmação: “Discordo totalmente” em azul, “Discordo” em laranja, “Indiferente” em verde, “Concordo parcialmente” em vermelho, “Concordo” em roxo, “Concordo totalmente” em rosa. FonteElaborada pelo Autor (2025).

Houve semelhança nas respostas entre os grupos analisados no formulário. Mesmo que quase 60% dos participantes tenham respondido que nunca ouviram falar do termo e 15% conheçam pouco sobre, houve respostas para consumo de produtos contendo leitelho pelo menos uma vez na semana, e a maior parte das opiniões indicou aceitação do leitelho ou de produtos à base dele, principalmente produtos de panificação e bebidas fermentadas, e os principais motivadores sendo curiosidade, sabor, e valor nutricional, que são favorecidos pelas propriedades do leitelho.

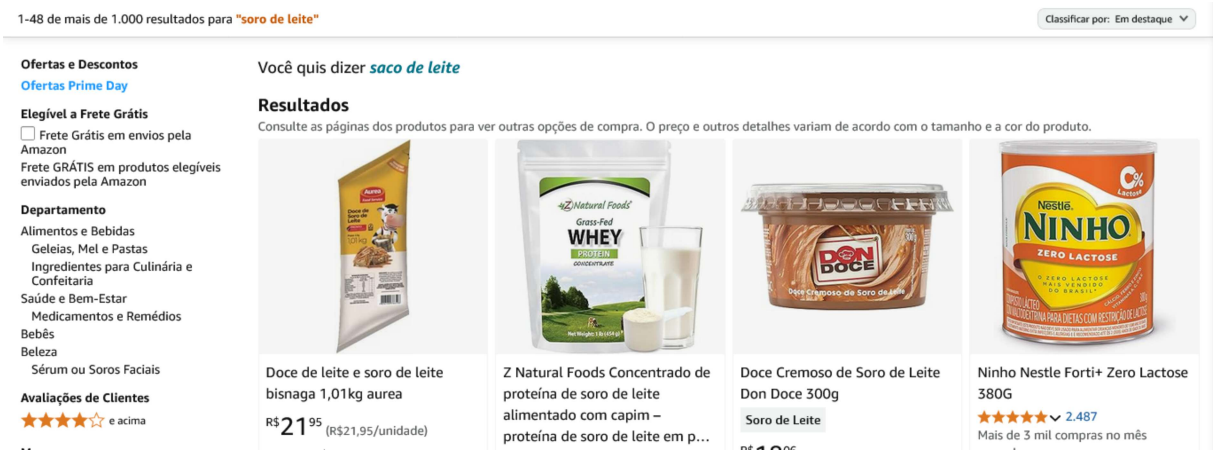
Entretanto, também houve grande ocorrência de “sabor ruim” e “textura ruim” como desmotivadores para experimentar os produtos, sendo as respostas mais escolhidas até para as pessoas que não são nada familiares com o termo leitelho, sendo necessário também informar a essas pessoas sobre os verdadeiros benefícios desse produto, não só no quesito sensorial, mas também para a saúde. Ainda assim, essas respostas negativas para “palavra, sentimento e/ou emoção” foram apenas uma pequena parcela do total.

O perfil de usuários com idade entre 25 e 44 anos e também com renda até R$5.000 foram os que mais responderam que há um interesse provável ou definitivo para experimentar alimentos contendo leitelho, que também foi um perfil que abrange grande parte dos participantes do formulário. Esse e futuros mapeamentos de perfil, em conjunto com pesquisas de mercado, ajudam a entender o mercado consumidor e qual é o público-alvo de possíveis novos produtos.

5.2 RASPAGEM DE DADOS

Buscando por produtos à base de leitelho pelas marcas mais predominantes, como Piracanjuba, Itambé ou Porto Alegre (MilkPoint, 2025), em *sites* com opção de avaliação do produto (exemplo, Amazon), os resultados foram inexpressivos não sendo possível coletar avaliações ou informações do perfil de quem avaliou. Por outro lado, para buscas por produtos à base de soro de leite (Figura 26), foi possível encontrar diferentes tipos de produtos de diferentes marcas, como doce de leite ou composto lácteo.

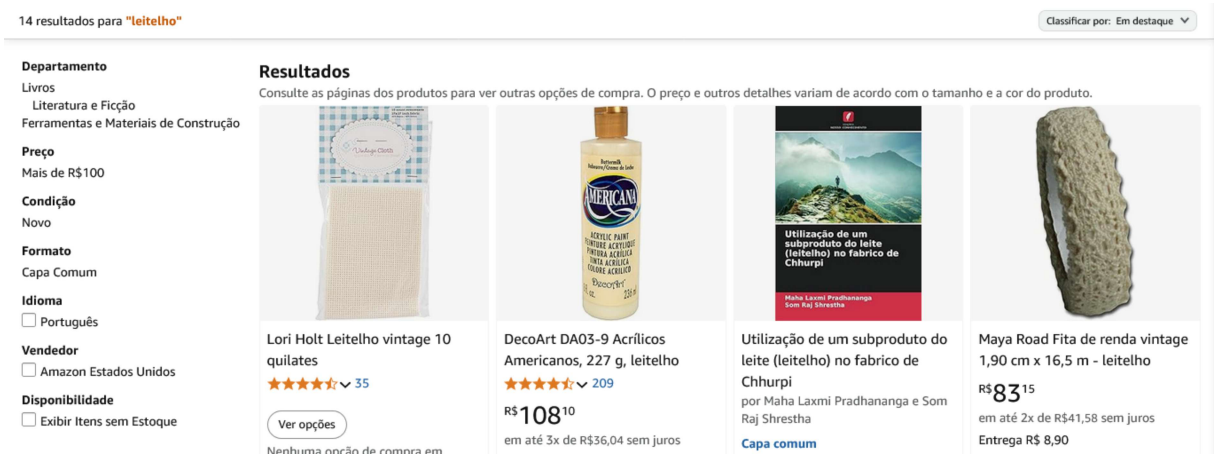
Figura 26 – Resultado de busca por “soro de leite” na Amazon.



Fonte: (Amazon, 2025).

No *site* da Amazon, as buscas retornaram menos resultados que no iHerb, seja buscando por “leitelho” ou por “*buttermilk*”, esses poucos eram misturas para panquecas ou para waffles, livros de receitas, garrafa para armazenar leite, e até objetos que têm a cor “leitelho” (como um amarelo pastel), como mostra a Figura 27.

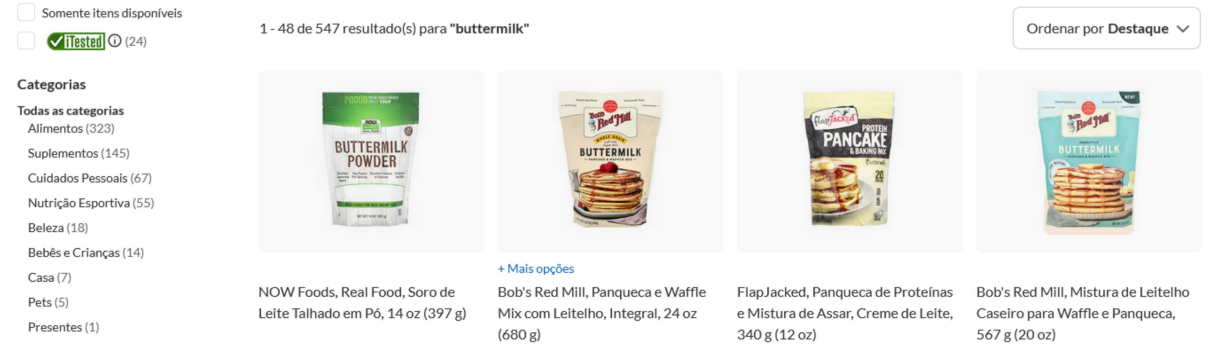
Figura 27 – Resultado de busca por “leitelho” na Amazon.



Fonte: (Amazon, 2025).

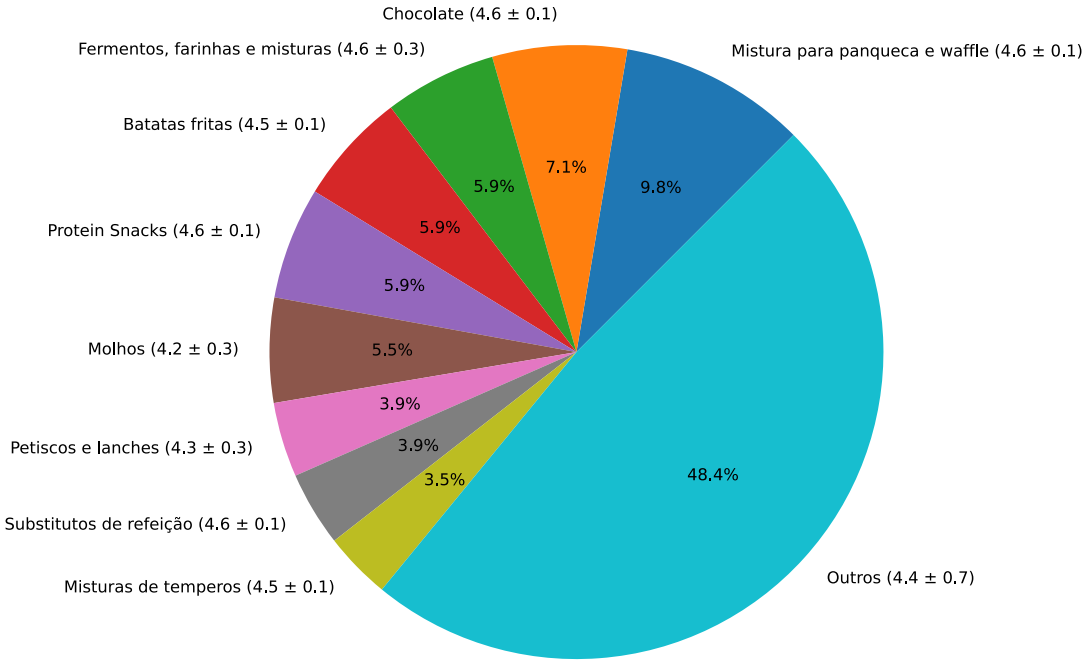
Já para o site iHerb, a variedade já é maior, tanto de marcas quanto de categorias (Figura 28). O gráfico de setores apresenta as 9 das 62 categorias mais presentes nesse site, agrupando as demais em “Outros”, acompanhadas de suas médias ± desvio padrão das avaliações deixadas pelos usuários (Figura 29). Foram coletados 254 produtos com códigos distintos, que vão de alimentos (como misturas para panificação ou molhos) até produtos de higiene de pessoas e também de animais. Essas categorias apresentam nota média acima de 4,0 (de máximo 5,0), e o maior desvio padrão de 1,1, sendo no agrupamento das categorias que não estão entre as mais avaliadas.

Figura 28 – Resultado de busca por “*buttermilk*” no site iHerb.



Fonte: (IHerb, 2025).

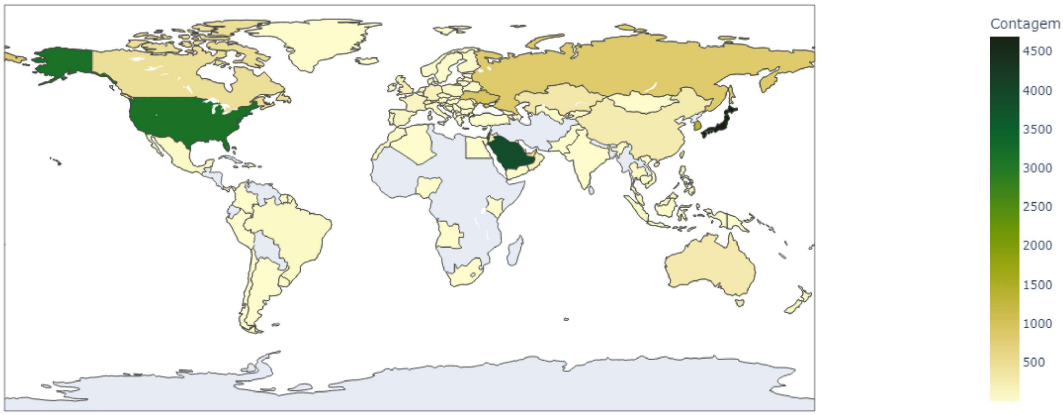
Figura 29 – Distribuição dos produtos com suas respectivas notas médias ± desvios padrões.



Fonte: Elaborada pelo Autor (2025).

Das 41.401 avaliações coletadas até o dia 18 de março de 2025, 28.415 são de usuários distintos. Esses usuários estão distribuídos entre 110 países, mostrados no mapa coroplético (Figura 30), dos quais os mais frequentes são do Japão (16,52%), Arábia Saudita (13,50%), Israel (12,68%), Hong Kong (12,67%), e dos Estados Unidos (11,00%), enquanto os 71 usuários brasileiros distintos representam apenas 0,24% desse total.

Figura 30 – Mapa coroplético da nacionalidade dos avaliadores distintos dos produtos comprados no iHerb.



Fonte: Elaborada pelo Autor (2025).

Após normalização e remoção das *stopwords* nas avaliações, restaram 507.756 das 1.098.589 palavras escritas (46,22% do total). Entre as 10 mais frequentes (Quadro 8), os adjetivos presentes ("delicioso", "bom", e "fácil") são todos positivos, similar ao caso das respostas do formulário apresentadas no Quadro 6 anteriormente. As palavras "sabor" e "saboroso" (similar a "delicioso") também estavam entre as mais respondidas no formulário.

Quadro 8 – Palavras mais citadas e suas respectivas frequências e porcentagens.

Palavra	**Frequência**	**Porcentagem**
sabor	12.360	2,43%
delicioso	5.640	1,11%
bom	4.828	0,95%
proteína	4.523	0,89%
ingrediente	4.416	0,87%
gosto	4.075	0,80%
qualidade	4.021	0,79%
fácil	3.968	0,78%
tamanho	3.526	0,69%
lanche	3.478	0,68%

Fonte: Elaborado pelo Autor (2025).

Também houve semelhança entre as polaridades dos adjetivos usados nas avaliações (Quadro 9) e as respostas do formulário (anteriormente no Quadro 7), sendo aqui nas avaliações 32.203 (52,22%) adjetivos positivos, 23.296 (46,67%) neutros, e apenas 1.781 (3,11%) negativos, havendo diferença estatística significativa entre todas elas (p-valor $\approx$ 0,00 em todos os casos). No contexto de compra, podemos ainda considerar que as palavras "ansioso", "viciante", e "viciado" foram escritas com um sentido positivo referente ao produto avaliado, como "ansioso pela entrega", ou "viciado em consumir esse produto", ou seja, reduzindo ainda mais a quantidade de adjetivos negativos nas avaliações nesse contexto. Também é preciso considerar que esses adjetivos não estão sendo ditos diretamente sobre o leitelho em si, mas sim sobre produtos contendo leitelho na composição.

Dos usuários distintos, apenas 3.286 (11,56%) possuíam alguma imagem de perfil diferente da padrão do *site*. Ainda que alguns usuários tenham imagem no perfil, essa não necessariamente continha uma pessoa, o que afetava a análise das características faciais. Das imagens coletadas, foi possível identificar face em 1.626 delas, referente a 49,48% dos usuários com imagens, mas apenas 5,72% do total de usuários distintos, sendo 772 (47,54%) identificados com sexo feminino e 852 (52,46%) identificados com masculino, não havendo diferença estatística significativa entre as proporções (p-valor $\approx$ 1,00).

Também não houve diferença estatística significativa entre as faixas etárias identificadas nas imagens de perfil (p-valor $\approx$ 1,00), diferente dos participantes do formulário. As distribuições individuais de sexo e de faixa etária são apresentadas no Quadro 10.

Quadro 9 – Classificação de palavras por polaridade e suas respectivas porcentagens.

Polaridade Negativa	**Polaridade Neutra**	**Polaridade Positiva**
Caro (15,72%)	Pouco (16,59%)	Bom* (20,80%)
Azedo* (10,98%)	Doce* (5,05%)	Boa (11,25%)
Ruim* (8,90%)	Grande (4,92%)	Delicioso (8,60%)
Estranho* (7,60%)	Alta (4,64%)	Excelente (6,63%)
Ansioso (6,61%)	Salgado (3,14%)	Melhor (4,90%)
Viciante (3,49%)	Picante (3,02%)	Perfeito (3,14%)
Viciado (2,65%)	Leve* (2,91%)	Rico (3,07%)
Pobre (2,55%)	Baixo (2,74%)	Crocante (2,95%)
Gorduroso* (2,50%)	Alto (1,77%)	Forte (2,94%)
Brega (1,82%)	Suave (1,65%)	Adequado (2,79%)

As palavras acompanhadas por "*" também estiveram entre as mais frequentes nas avaliações coletadas na raspagem de dados, apresentadas posteriormente no Quadro 6. Fonte: Elaborado pelo Autor (2025).

Não foram estimados rostos com 55 anos ou mais, além de duas imagens estimadas com menos de 18 anos, um do sexo masculino e outro do feminino, sendo excluídos das análises posteriores, tanto por representarem uma proporção pequena entre as outras faixas quanto por não terem participação dessa faixa etária no formulário da UFV.

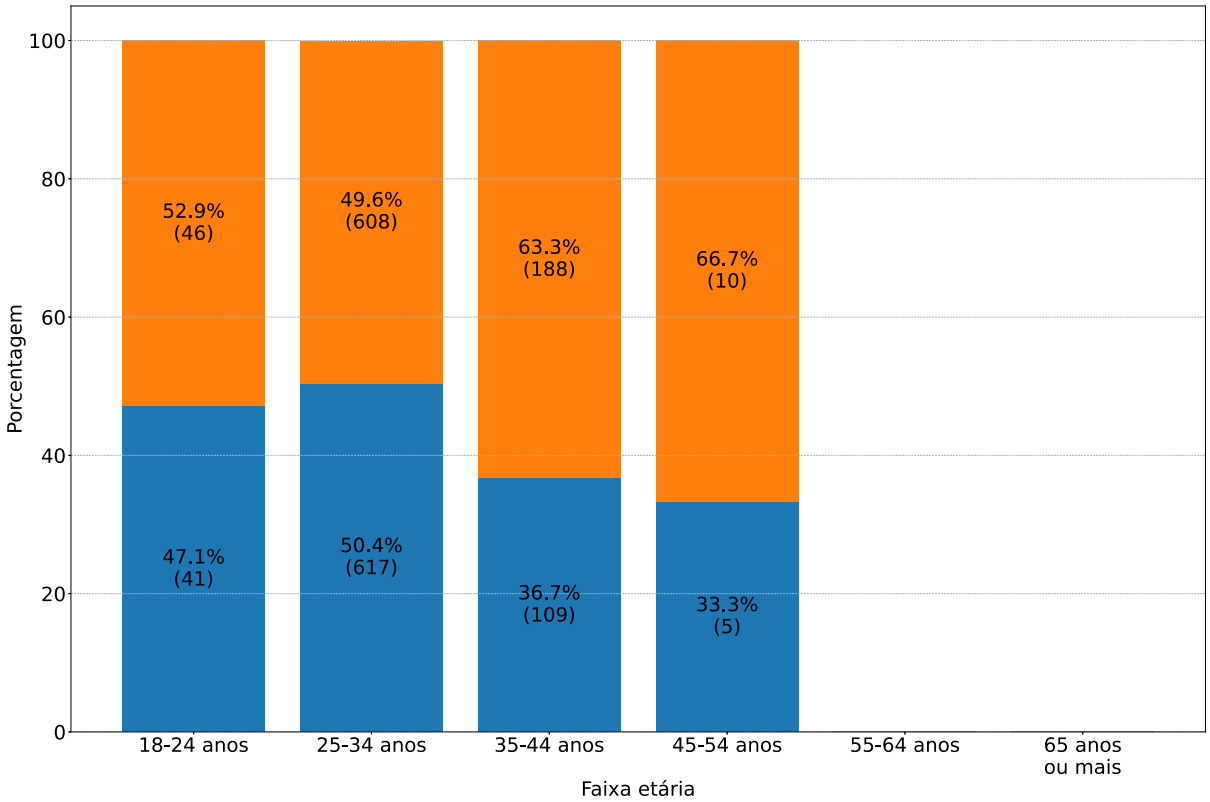
Quadro 10 – Distribuição das faixas etárias dos respondentes do formulário.

Faixa Etária	**Contagem**	**Proporção**
18-24 anos	87	5,35%
25-34 anos	1225	75,43%
35-44 anos	297	18,28%
45-54 anos	15	0,92%
55-64 anos	0	0%
65 anos ou mais	0	0%

Fonte: Elaborado pelo Autor (2025).

A Figura 31 traz a distribuição dos gêneros por faixa etária. Houve diferença estatística significativa entre as faixas etárias ao agrupar por sexo (p-valor $\approx 0{,}00$), com o sexo feminino mais presente entre 25 e 34 anos, essa sendo a única faixa onde o sexo feminino foi maioria, e também menos presente na faixa seguinte. Apenas 9 usuários brasileiros tiveram as características faciais estimadas, sendo 3 do sexo masculino e 2 do feminino entre 35 e 44 anos, e 6 do sexo feminino entre 25 e 34 anos.

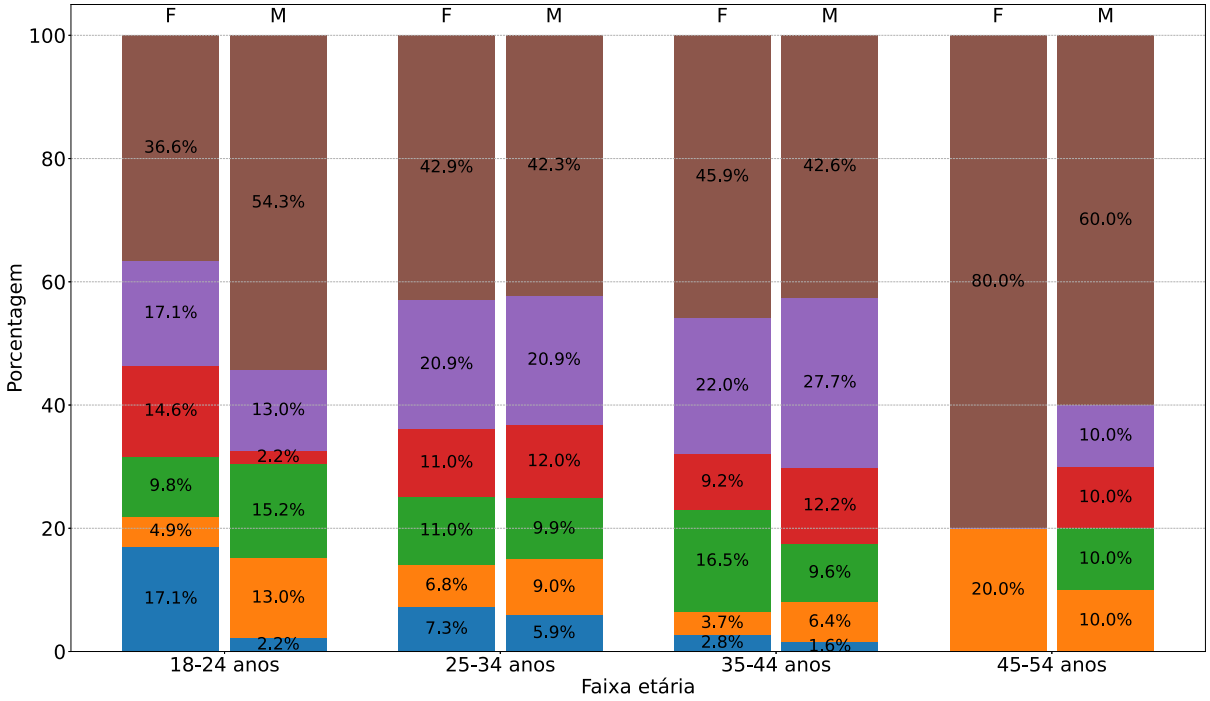
Figura 31 – Distribuição de sexo dos usuários por faixa etária



Valores percentuais e absolutos individuais em cada segmento, sem registros para “55-64 anos” e “65-anos ou mais”, por gênero: “Feminino” em azul, “Masculino” em laranja. Fonte: Elaborada pelo Autor (2025).

Selecionando as 5 categorias com mais avaliações e agrupando as demais como “Outros”, é possível observar que essas mais avaliadas representam cerca de 50% do total em todas as faixas etárias (Figura 32). Não houve avaliação de usuários entre 45 e 54 anos em produtos da categoria “Lanches, Barras e Alimentos para Comer com as Mãos”, categoria que foi mais avaliada pelo sexo feminino de 18 a 24 anos do que nas outras faixas, com diferença estatística significativa (p-valor $\approx 0{,}04$).

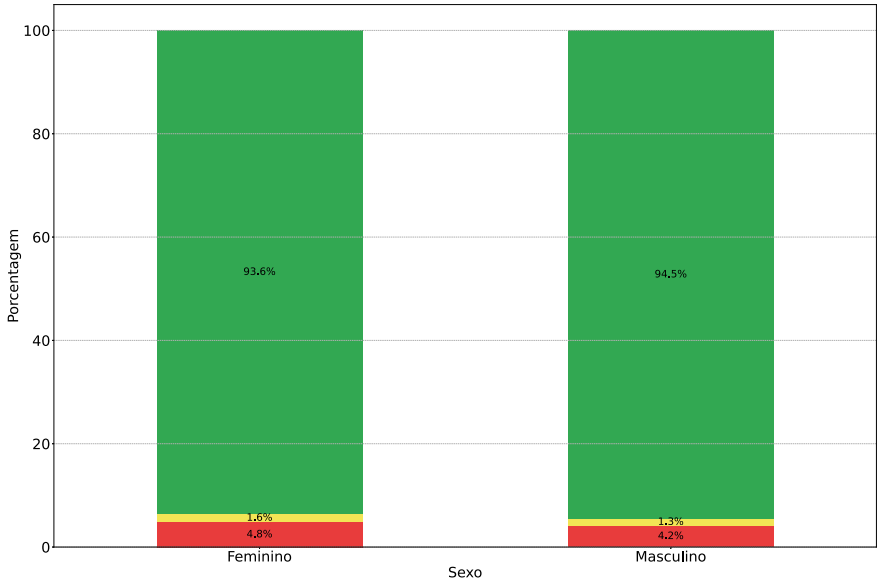
Figura 32 – Distribuição de categorias avaliadas pelos usuários por sexo em cada faixa etária



Para cada faixa etária, separados nas barras F e M para os gêneros feminino e masculino, respectivamente, os segmentos apresentam os valores absolutos para a frequência de consumo dos produtos: “Lanches, Barras e Alimentos para comer com as mãos” em azul, “Petiscos e lanches” em laranja, “Misturas para temperos” em verde, “Batatas fritas” em vermelho, “Mistura para panqueca e waffle” em roxo, “Outros” em marrom. Fonte: Elaborada pelo Autor (2025).

A análise de sentimentos das avaliações coletadas (Figura 33) está de acordo com as notas numéricas deixadas pelos usuários e também segue uma distribuição ainda mais positivista que na análise das “palavras, sentimentos ou emoções” respondidas no formulário (anteriormente na Figura 16), com os adjetivos positivos estando acima de 90% para os dois sexos, seguidos por adjetivos negativos e, por último, os adjetivos neutros. Houve diferença significativa entre as três polaridades (p-valores $\approx$ 0,00 nos três casos), mas não houve ao testar também entre os sexos (p-valor $\approx$ 0,61). Contudo, ainda que as polaridades sejam mais positivas nos produtos coletados do que nas respostas do formulário, essas polaridades não necessariamente se referem ao leitelho em si, mas sim aos produtos que contêm leitelho.

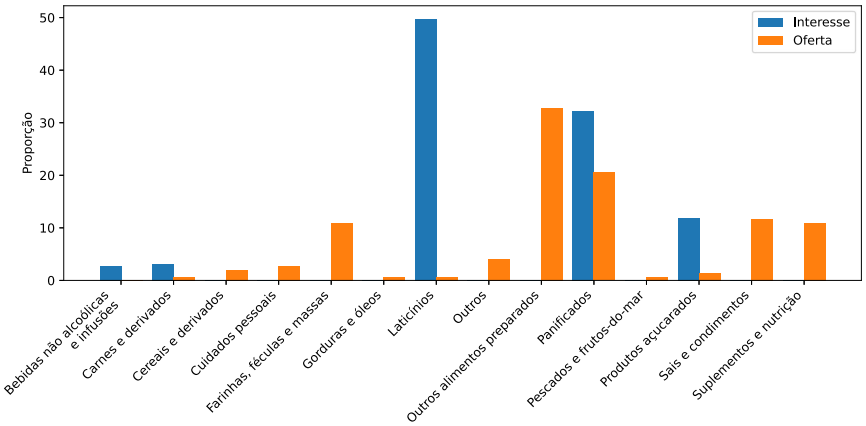
Figura 33 – Distribuição das polaridades das avaliações coletadas.



Para cada sexo, os valores percentuais das polaridades são apresentados, sendo verde para positivas, amarelo para neutras, vermelho para negativas. Fonte: Elaborada pelo Autor (2025).

Os produtos coletados apresentam avaliações de usuários de vários países, não apenas do Brasil, e é possível observar que as avaliações, no geral, são muito positivas e também que algumas categorias desses produtos estão alinhadas com aquelas que os respondentes do formulário marcaram como interesse em experimentar, mas que há uma grande oportunidade que pode ser explorada, como o interesse por laticínios, bebidas não alcoólicas e infusões, panificados, carnes e derivados, e produtos açucarados (Figura 34).

Figura 34 – Distribuição de categorias contendo leitelho de interesse e de oferta.



Para cada categoria, a barra azul (à esquerda) representa a proporção de produtos de interesse que foram respondidos no formulário, e a barra laranja (à direita) representa a proporção de produtos coletados na raspagem de dados. Fonte: Elaborada pelo Autor (2025).

O Quadro 11 traz as categorias dos produtos que foram avaliados por usuários brasileiros, com média acima de 4,0 em todas elas. Houve maior presença de produtos panificados, sais e condimentos, e outros alimentos preparados, que foram as categorias com mais produtos disponíveis, enquanto apenas um laticínio foi avaliado, que estava entre as menores categorias, podendo estar limitado por ser um produto que requer refrigeração adequada.

Quadro 11 – Distribuição das categorias de produtos vendidos com estatísticas.

Categoria	**Nota média**	**Desvio padrão**	**n**	**Proporção**
Panificados	4,51	0,13	25	35,21%
Sais e condimentos	4,52	0,13	25	35,21%
Outros alimentos preparados	4,41	0,10	10	14,08%
Farinhas, féculas e massas	4,60	0,00	4	5,63%
Cuidados pessoais	4,17	0,06	3	4,23%
Suplementos e nutrição	4,60	0,00	2	2,82%
Laticínios	4,70	—	1	1,41%
Pescados e frutos-do-mar	4,70	—	1	1,41%

Fonte: Elaborado pelo Autor (2025).

6 DISCUSSÃO

Os produtos que as pessoas responderam no formulário como interesse em experimentar estão de acordo com as categorias de produtos listados nos sites de compras que foram coletados através da raspagem de dados, como principalmente os panificados, além de semelhanças nos principais adjetivos usados nas avaliações dos produtos comprados com as "palavras, sentimentos e/ou emoções" respondidas pelos participantes do formulário.

Comparando os produtos com leitelho coletados e os produtos de interesse que foram respondidos no formulário, é considerável que há um mercado a ser explorado, principalmente levando em conta que existem pouquíssimos produtos nacionais produzidos com leitelho como substituto do leite. Os usuários brasileiros avaliaram mais na categoria de panificação, que é a que mais traz variedades, enquanto os laticínios, que mais geraram interesse no formulário, foram avaliados apenas uma vez. Mas, vale ressaltar que as pessoas podem comprar e não avaliar, o que pode alterar essas contagens, ou que ainda não compram por não terem conhecimento do que é o leitelho e quais são seus benefícios.

Para as avaliações dos produtos coletados pela raspagem de dados, as avaliações foram positivas em sua grande maioria, e ainda são produtos que se encaixam nas categorias que os participantes do formulário responderam que têm interesse em experimentar, indicando que há um mercado a ser explorado no país, mesmo com os usuários brasileiros nesse site sendo apenas uma pequena parcela dos que avaliaram tais produtos. Entretanto, é importante levar em conta que existem algumas limitações ao considerar apenas as avaliações coletadas, já que as pessoas podem comprar sem avaliar o produto, ou seja, pode ser que outras categorias sejam mais compradas em vez dessas que são mais avaliadas, ou que haja ainda mais brasileiros comprando esses produtos.

Não foi possível confirmar se os usuários dos dados raspados atualizam suas imagens de perfil, sendo um limitador para a extração de características faciais, estimando faixas etárias que não condizem totalmente com a pessoa que avaliou o produto. Também houve casos de imagens de perfil que não eram com apenas uma pessoa, ou até com nenhuma pessoa, não sendo possível estimar as características dessas imagens.

Uma análise sensorial realizada por (Teixeira *et al.*, 2020) apresenta baixa aceitação de uma bebida achocolatada à base de leite e leitelho, principalmente entre os participantes mais jovens, com falta de doçura (dulçor) e sabor de chocolate sendo os motivos da avaliação negativa. Já para sorvete à base de leitelho (Ramos *et al.*, 2021), a amostra com 100% de leitelho teve melhores avaliações para sabor, cor, e textura, além de melhor derretimento e maior capacidade para incorporação de ar do que as amostras tradicional e com 50% de leitelho na análise sensorial.

Apesar dos estudos, ao buscarmos por "sorvete à base de leitelho" na *internet*, não temos resultados de produtos das principais marcas à venda, como Nestlé (MilkPoint,

2025), apenas da marca “Cremoso” (marca de sorvete e picolé), que lista o leitelho no ingrediente de seus sorvetes dentro de “Leite Integral e/ou Composto Lácteo”, como mostra um exemplo na Figura 35.

Figura 35 – Lista de ingredientes de um sorvete da marca “Cremoso”.

Flocos

INGREDIENTES: Leite Integral e/ou Composto Lácteo ((Soro de Leite, Gordura Vegetal, Concentrado Proteico de Soro de Leite, Maltodextrina, Soro de Leite, Açúcar, Leite em Pó Desnatado e Integral, Leitelho, Emulsificantes: Ésteres de Mono e Diglicerídeos de Ácidos Graxos com Ácido Lático (INS 472b), Estearoil Lactato de Sódio (INS 481i) e Mono e Diglicerídeos de Ácidos Graxos (INS 471), Estabilizantes: Carboximetilcelulose (INS 466) e Celulose Microcristalina (INS 460i), Aromatizantes e Regulador de Acidez: Bicarbonato De Sódio (INS 500ii)), Açúcar, Cobertura de Chocolate (Açúcar, Gordura Vegetal Modificada, Cacau em Pó Processado com Álcali, Pasta De Cacau, Emulsificante: Lecitina De Soja, Emulsificante: Ésteres De Ácido Ricinoléico

Fonte: (Cremoso, 2024).

Já é comum no Brasil a venda de produtos lácteos tendo o leite integral parcialmente substituído pelo soro do leite e que também possuem embalagens similares ao que seria o produto “original” (prática que pode ser caracterizada como violação do *trade dress*). Essa prática, ainda que questionável, é comum em produtos com o soro de leite, como nos compostos lácteos em pó ou misturas lácteas condensadas (Braz, 2022), novamente indicando que o leitelho também tem mais espaço para ser utilizado como ingrediente na indústria alimentícia.

Ainda que seja possível encontrar produtos industrializados nacionais com leitelho, como a “Bebida láctea Frutirol” da Tirol (Figura 36) ou “Composto lácteo profissional” da Itambé (este não sendo encontrado no *site* da sua marca, mas em revendedores), além do sorvete já mencionado, são poucos e em lojas em que não foi possível coletar as avaliações ou informações de quem avaliou, limitando as análises de perfil e de sentimentos.

Figura 36 – Lista de ingredientes de uma bebida láctea da marca “Tirol”.

INGREDIENTES

Soro de leite e/ou soro de leite reconstituído, leitelho e/ou leitelho em pó reconstituído, leite pasteurizado e/ou leite em pó reconstituído, açúcar, preparado de coco (água, açúcar, leite de coco, amido modificado, aromatizante, acidulante ácido lático, conservante sorbato de potássio e corante inorgânico dióxido de titânio), amido modificado, fermento lácteo e espessante pectina cítrica. CONTÉM LACTOSE. ALÉRGICOS: CONTÉM LEITE E DERIVADOS. ESTE PRODUTO NÃO É LEITE. NÃO CONTÉM GLÚTEN.

BEBIDA LÁCTEA NÃO É IOGURTE. Mantenha resfriado de 1°C a 10°C. Após aberto consumir em até 3 dias.

Fonte: (Tirol, 2025).

A curiosidade foi uma resposta muito comum entre os participantes do formulário quando perguntados sobre o motivo que os levaria a experimentar produtos contendo leitelho, junto com sabor e preço baixo. Já como motivo para não experimentar, as principais respostas foram sabor ruim e textura ruim. Essas respostas foram as principais até em quem não tem conhecimento sobre o termo.

Considerando que não houve falta de entendimento ao responder essas perguntas, vemos que é necessário uma forma de informar às pessoas sobre o que é o leitelho e como ele tem suas propriedades similares ao leite, além de ele já ser usado em receitas, como já citado, nas panificações ou bebidas lácteas. Os meios de informação que os participantes responderam no formulário, os principais sendo *sites* na internet e redes sociais, também indicam onde são os principais pontos que podem ser explorados a fim de informá-los melhor sobre produtos e benefícios desse coproduto.

7 CONCLUSÃO

Apesar de ser possível encontrar em sites da *internet* diversos produtos que usam leitelho na sua composição, foram poucos produtos nacionais com leitelho sendo produzidos em grande escala ou importados e vendidos nos sites de vendas avaliados nesse trabalho. Além disso, não foi possível coletar todas as avaliações ou informações detalhadas de quem comprou, apenas idade e sexo estimados, e país de origem. No entanto, a loja iHerb apresentou grande variedade de produtos com leitelho, bem avaliados, e esses dados contribuíram para compreender melhor o perfil de consumo e o interesse do público por esse tipo de produto.

A análise exploratória dos dados coletados permitiu identificar padrões de consumo, percepções e lacunas de conhecimento sobre o produto. Os resultados do formulário mostram como o leitelho é pouco consumido e até pouco conhecido pela população brasileira. Esses dados contribuem para a indústria de laticínios ao indicar como apresentar e posicionar o leitelho no mercado nacional.

Ao combinar as opiniões coletadas dos variados produtos no mercado internacional com os produtos que despertam o interesse da população brasileira, como pães e bebidas lácteas, identificou-se que o leitelho tem um potencial mercadológico forte. As motivações para experimentar e os tipos de produtos preferidos mostraram-se semelhantes entre diferentes perfis de pessoas, o que reforça a convergência entre as avaliações internacionais e as percepções locais.

A análise das variedades de produtos e do perfil de seus compradores, tanto nacionais quanto internacionais, mostrou que a utilização do leitelho em larga escala pode gerar maior lucratividade para a indústria de laticínios. Além disso, traz benefícios à saúde de quem o consome e reduz o desperdício desse coproduto.

Trabalhos futuros devem aprimorar a abordagem da coleta de produtos nos *sites* e coletar outros produtos avaliados pelos usuários, visando mapear o perfil de consumo deles para um melhor entendimento do público que pode estar interessado na compra de leitelho e produtos à base dele. Também deve-se entender como as grandes empresas utilizam o leitelho que produzem.

REFERÊNCIAS

ABDI, Hervé; WILLIAMS, Lynne J. Principal component analysis. **WIREs Computational Statistics**, v. 2, n. 4, p. 433–459, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101. ISSN 1939-0068. DOI: 10.1002/wics.101. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101. Acesso em: 6 mar. 2025.

ALPAYDIN, Ethem. **Introduction to Machine Learning**. 4. ed. Cambridge, MA, USA: MIT Press, mar. 2020. (Adaptive Computation and Machine Learning series). ISBN 978-0-262-04379-3.

AMAZON. **Amazon.com.br | Tudo pra você, de A a Z.** Disponível em: https://www.amazon.com.br/. Acesso em: 11 jul. 2025.

ANUÁRIO Leite 2023: leite baixo carbono. - Portal Embrapa. Disponível em: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1154264/anuario-leite-2023-leite-baixo-carbono. Acesso em: 3 abr. 2025.

BARBOSA, Jardeson Leandro Nascimento *et al.* Introdução ao Processamento de Linguagem Natural usando Python, jun. 2017.

BARDIN, Laurence. **Análise de conteúdo**. [*S. l.*]: Edições 70, 24 mar. 2011. ISBN 978-85-62938-04-7.

BIRD, Steven. Natural Language Processing with Python.

BIRD, Steven; LOPER, Edward. NLTK: The Natural Language Toolkit. *In:* PROCEEDINGS of the ACL Interactive Poster and Demonstration Sessions. Barcelona, Spain: Association for Computational Linguistics, jul. 2004. p. 214–217. Disponível em: https://aclanthology.org/P04-3031. Acesso em: 21 out. 2024.

BRASIL, Ministerio da Agricultura e Pecuária. **Mapa do Leite**. Ministério da Agricultura e Pecuária. 26 mar. 2025. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/producao-animal/mapa-do-leite/mapa-do-leite. Acesso em: 2 abr. 2025.

BRAZ, Kecieli Martins. ESTUDO DO GRAU DE SIMILARIDADE ENTRE EMBALAGENS DE PRODUTOS LÁCTEOS, 2022.

CAMBRIA, Erik *et al.* SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. *In:* COLING 2016. **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Edição: Yuji Matsumoto e Rashmi Prasad. Osaka, Japan: The COLING 2016 Organizing

Committee, dez. 2016. p. 2666–2677. Disponível em: https://aclanthology.org/C16-1251. Acesso em: 19 set. 2024.

CAMPOS, Claudinei José Gomes. Método de análise de conteúdo: ferramenta para a análise de dados qualitativos no campo da saúde. **Revista Brasileira de Enfermagem**, v. 57, p. 611–614, out. 2004. Publisher: Associação Brasileira de Enfermagem. ISSN 0034-7167, 1984-0446. DOI: https://doi.org/10.1590/S0034-71672004000500019. Disponível em: https://www.scielo.br/j/reben/a/wBbjs9fZBDrM3c3x4bDd3rc/. Acesso em: 13 mar. 2025.

CAPPELLE, Mônica Carvalho Alves; MELO, Marlene Catarina de Oliveira Lopes; GONÇALVES, Carlos Alberto. Análise de conteúdo e análise de discurso nas ciências sociais. **Organizações Rurais & Agroindustriais**, v. 5, n. 1, 2003. Number: 1. ISSN 2238-6890. Disponível em: https://www.revista.dae.ufla.br/index.php/ora/article/view/251. Acesso em: 28 mar. 2025.

CARDOSO, Márcia Regina Gonçalves; OLIVEIRA, Guilherme Saramago de; GHELLI, Kelma Gomes Mendonça. ANÁLISE DE CONTEÚDO: UMA METODOLOGIA DE PESQUISA QUALITATIVA. **Cadernos da FUCAMP**, v. 20, n. 43, 25 mar. 2021. Number: 43. ISSN 2236-9929. Disponível em: https://revistas.fucamp.edu.br/index.php/cadernos/article/view/2347. Acesso em: 14 mar. 2025.

COSTA, Marcela de Rezende. **Obtenção de ingrediente lacteo enriquecido em lipideos polares a partir de leitelho de soro**. [*S. l.*], 2008. Disponível em: https://repositorio.unicamp.br/acervo/detalhe/435882. Acesso em: 18 fev. 2025.

CREMOSO. **Flocos**. CREMOSO - O melhor sorvete! 2024. Disponível em: https://www.cremoso.com.br/_loja_/p/4794136/flocos. Acesso em: 16 out. 2024.

DEWETTINCK, Koen *et al.* Nutritional and technological aspects of milk fat globule membrane material. **International Dairy Journal**, v. 18, n. 5, p. 436–457, maio 2008. ISSN 09586946. DOI: 10.1016/j.idairyj.2007.10.014. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0958694607002336. Acesso em: 18 fev. 2025.

ETEMAD, Kamran; CHELLAPPA, Rama. Discriminant analysis for recognition of human face images. **Journal of the Optical Society of America A**, v. 14, n. 8, p. 1724, 1 ago. 1997. ISSN 1084-7529, 1520-8532. DOI: 10.1364/JOSAA.14.001724. Disponível em: https://opg.optica.org/abstract.cfm?URI=josaa-14-8-1724. Acesso em: 11 maio 2025.

FACELI, Katti. **Inteligência artificial : uma abordagem de aprendizado de máquina**. [*S. l.*]: LCT, 2021. Accepted: 2022-04-20T20:34:56Z. ISBN 978-85-216-3734-9. Disponível em: http://bibliotecadigital.tse.jus.br/xmlui/handle/bdtse/10079. Acesso em: 9 fev. 2025.

FERNANDES, Fernando Timoteo; CHIAVEGATTO FILHO, Alexandre Dias Porto. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. **Revista Brasileira de Saúde Ocupacional**, v. 44, e13, 4 nov. 2019. Publisher: Fundação Jorge Duprat Figueiredo de Segurança e Medicina do Trabalho - FUNDACENTRO. ISSN 0303-7657, 2317-6369. DOI: https://doi.org/10.1590/2317-6369000019418. Disponível em: https://www.scielo.br/j/rbso/a/NgxW5qxzQWhcD4KrTHLxxGG/. Acesso em: 4 mar. 2025.

GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. [*S. l.*]: O'Reilly Media, Inc., 2019. Google-Books-ID: HHetDwAAQBAJ. ISBN 978-1-4920-3261-8.

HONGYU, Kuang; SANDANIELO, Vera Lúcia Martins;
JUNIOR, Gilmar Jorge de Oliveira. Análise de Componentes Principais: Resumo Teórico, Aplicação e Interpretação. **E&S Engineering and Science**, v. 5, n. 1, p. 83–90, 29 jun. 2016. ISSN 2358-5390. DOI: 10.18607/ES201653398. Disponível em: https://periodicoscientificos.ufmt.br/ojs/index.php/eng/article/view/3398.

HTTPS://CIAGRO.INSTITUTOIDV.ORG/CIAGRO/UPLOADS/483.PDF. Disponível em: https://ciagro.institutoidv.org/ciagro/uploads/483.pdf. Acesso em: 18 abr. 2024.

HUNTER, John D. Matplotlib: A 2D Graphics Environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, maio 2007. Conference Name: Computing in Science & Engineering. ISSN 1558-366X. DOI: 10.1109/MCSE.2007.55. Disponível em: https://ieeexplore.ieee.org/document/4160265. Acesso em: 30 out. 2024.

IHERB. **iHerb | vitaminas, suplementos e produtos naturais para saúde**. Disponível em: https://br.iherb.com/. Acesso em: 11 jul. 2025.

IZBICKI, Rafael. **Aprendizado de máquina: uma abordagem estatística**. Em colaboração com Tiago Mendonça dos Santos. [*S. l.*]: Rafael Izbicki, 29 abr. 2020. ISBN 9786500024104.

AL-JAHANI, Amani H. Effect of Buttermilk on the Physicochemical, Rheological, and Sensory Qualities of Pan and Pita Bread. **International Journal of Food Science**, v. 2017, n. 1, p. 2054252, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2017/2054252. ISSN 2314-5765. DOI: 10.1155/2017/2054252. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/2054252. Acesso em: 21 mar. 2025.

JUNCZYS-DOWMUNT, Marcin *et al.* Marian: Fast Neural Machine Translation in C++. *In:* LIU, Fei; SOLORIO, Thamar (ed.). **Proceedings of ACL 2018, System Demonstrations**. Melbourne, Australia: Association for Computational Linguistics, jul.

2018. p. 116–121. DOI: 10.18653/v1/P18-4020. Disponível em: https://aclanthology.org/P18-4020. Acesso em: 5 jun. 2024.

LIMA, Jorge Ávila de. Por uma Análise de Conteúdo Mais Fiável. **Revista Portuguesa de Pedagogia**, p. 7–29, 2013. ISSN 1647-8614. DOI: 10.14195/1647-8614_47-1_1. Disponível em: https://impactum-journals.uc.pt/rppedagogia/article/view/1647-8614_47-1_1. Acesso em: 13 mar. 2025.

LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, v. 35, p. 85–94, 19 abr. 2021. Publisher: Instituto de Estudos Avançados da Universidade de São Paulo. ISSN 0103-4014, 1806-9592. DOI: https://doi.org/10.1590/s0103-4014.2021.35101.007. Disponível em: https://www.scielo.br/j/ea/a/wXBdv8yHBV9xHz8qG5RCgZd/?lang=pt&format=html. Acesso em: 4 mar. 2025.

MACHADO, Ezequiel Luiz; RAMOS, Gaspar Dias Monteiro; ANTUNES, Veridiana de Carvalho. O leitelho e sua utilização pela indústria de alimentos. **Revista do Instituto de Laticínios Cândido Tostes**, v. 77, n. 1, p. 43–54, 25 abr. 2022. Number: 1. ISSN 2238-6416. DOI: 10.14295/2238-6416.v77i1.873. Disponível em: https://www.revistadoilct.com.br/rilct/article/view/873. Acesso em: 20 mar. 2025.

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. DOI: 10.25080/Majora-92bf1922-00a.

MILKPOINT. **Histórico regulatório da manteiga: o que é manteiga extra, de primeira ou segunda qualidade?** MilkPoint. 13 set. 2023. Disponível em: https://www.milkpoint.com.br/colunas/lipaufv/historico-regulatorio-da-manteiga-o-que-e-manteiga-extra-de-primeira-ou-segunda-qualidade-234979/. Acesso em: 1 abr. 2025.

MILKPOINT. **Leitelho: a riqueza que vem da manteiga**. MilkPoint. Disponível em: https://www.milkpoint.com.br/colunas/cpa-ufg/leitelho-a-riqueza-que-vem-da-manteiga-233162/. Acesso em: 18 abr. 2024.

MILKPOINT. **Ranking dos Maiores Laticínios do Brasil 2024: volume de leite de produtores segue em crescimento**. MilkPoint. 24 mar. 2025. Disponível em: https://www.milkpoint.com.br/noticias-e-mercado/giro-noticias/ranking-dos-maiores-laticinios-do-brasil-2024-volume-de-leite-de-produtores-segue-em-crescimento-238278/. Acesso em: 25 mar. 2025.

MITCHELL, Ryan. **Web Scraping with Python**. [*S. l.*]: "O'Reilly Media, Inc.", 14 fev. 2024. 352 p. Google-Books-ID: ycf0EAAAQBAJ. ISBN 978-1-09-814531-6.

MONTEIRO, F. C. *et al.* Manteiga com adição de ervas finas: Inovação no setor de mercado lácteo. **Revista ESPACIOS | Vol. 35 (Nº 13) Año 2014**, 11 dez. 2014. Disponível em: https://www.revistaespacios.com/a14v35n13/14351308.html. Acesso em: 20 mar. 2024.

MUELLER, Andreas. **A Wordcloud in Python**. A Wordcloud in Python. 6 nov. 2012. Disponível em: https://peekaboo-vision.blogspot.com/2012/11/a-wordcloud-in-python.html. Acesso em: 30 out. 2024.

NAOKI, Iana. **PyMySQL documentation — PyMySQL 0.7.2 documentation**. Disponível em: https://pymysql.readthedocs.io/en/latest/. Acesso em: 31 out. 2024.

NASTESKI, Vladimir. An overview of the supervised machine learning methods. **HORIZONS.B**, v. 4, p. 51–62, 15 dez. 2017. ISSN 18578578, 18579892. DOI: 10.20544/HORIZONS.B.04.1.17.P05. Disponível em: http://uklo.edu.mk/filemanager/HORIZONTI%202017/Serija%20B%20br.%204/6.An%20overview%20of%20the%20supervised.pdf. Acesso em: 11 maio 2025.

NOGUEIRA, Thallys; MONTEIRO, Anna *et al.* Mineração de dados em tweets para análise do consumo de lácteos no Brasil. **The Journal of Engineering and Exact Sciences**, v. 8, 14863–01a, 1 dez. 2022. DOI: 10.18540/jcecvl8iss10pp14863-01a.

NOGUEIRA, Thallys; SIQUEIRA, Kennya; CAPRILES, Priscila. Construction of a training dataset for a sentiment analysis model of dairy products tweets in Brazil. **Social Network Analysis and Mining**, v. 14, 15 abr. 2024. DOI: 10.1007/s13278-024-01254-5.

NOGUEIRA, Thallys da Silva. **Observatório do consumidor: uma ferramenta de mineração de dados de redes sociais para avaliação de tendências de consumo de derivados lácteos no Brasil**. Mar. 2025. Tese (Doutorado) – Universidade Federal de Juiz de Fora (UFJF). Accepted: 2025-05-29T11:18:44Z. Disponível em: https://repositorio.ufjf.br/jspui/handle/ufjf/18795.

OLIVEIRA, Débora F. de; BRAVO, Claudia E. C.; TONIAL, Ivane B. SORO DE LEITE: UM SUBPRODUTO VALIOSO. **Revista do Instituto de Laticínios Cândido Tostes**, v. 67, n. 385, p. 64–71, 2012. Number: 385. ISSN 2238-6416. DOI: 10.5935/2238-6416.20120025. Disponível em: https://www.revistadoilct.com.br/rilct/article/view/215. Acesso em: 11 jul. 2025.

OLIVEIRA, Johnatan Santos De. CROSS-DOMAIN DEEP FACE MATCHING FOR BANKING SECURITY SYSTEMS, maio 2018.

PALMER, David D. Text Preprocessing. *In:* INDURKHYA, Nitin; DAMERAU, Fred J. (ed.). **Handbook of Natural Language Processing**. 2. ed. [*S. l.*]: Chapman e Hall/CRC, 2010. Num Pages: 22. ISBN 978-0-429-14920-7.

PEREIRA, Alessandro Campos. Tipos de leitelho gerado pelas indústrias de laticínios fiscalizadas pelo serviço de inspeção federal e sua destinação. **Revista do Instituto de Laticínios Cândido Tostes**, v. 77, n. 2, p. 103–110, 1 dez. 2022. Number: 2. ISSN 2238-6416. DOI: 10.14295/2238-6416.v77i2.893. Disponível em: https://www.revistadoilct.com.br/rilct/article/view/893. Acesso em: 18 abr. 2024.

PEREIRA, Amanda Lucas. **APRENDIZADO SEMI E AUTO-SUPERVISIONADO APLICADO À CLASSIFICAÇÃO MULTI-LABEL DE IMAGENS DE INSPEÇÕES SUBMARINAS**. 10 mar. 2023. MESTRE EM ENGENHARIA ELÉTRICA – PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO, Rio de Janeiro, Brazil. DOI: 10.17771/PUCRio.acad.63187. Disponível em: http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=63187@1. Acesso em: 11 maio 2025.

PERKTOLD, Josef; SEABOLD, Skipper. **statsmodels/statsmodels: Release 0.14.2**. [*S. l.*]: Zenodo, 17 abr. 2024. DOI: 10.5281/ZENODO.593847. Disponível em: https://zenodo.org/doi/10.5281/zenodo.593847. Acesso em: 1 nov. 2024.

RAINA, Rajat *et al.* Self-taught learning: transfer learning from unlabeled data. *In:* PROCEEDINGS of the 24th international conference on Machine learning. New York, NY, USA: Association for Computing Machinery, 20 jun. 2007. (ICML '07), p. 759–766. ISBN 978-1-59593-793-3. DOI: 10.1145/1273496.1273592. Disponível em: https://doi.org/10.1145/1273496.1273592. Acesso em: 6 mar. 2025.

RAMOS, Isabella *et al.* Desenvolvimento de sorvete com adição de leitelho. **Brazilian Journal of Food Technology**, v. 24, e2020237, 23 jul. 2021. Publisher: Instituto de Tecnologia de Alimentos - ITAL. ISSN 1981-6723. DOI: 10.1590/1981-6723.23720. Disponível em: https://www.scielo.br/j/bjft/a/YLh4fjsRXCQNNkDbmCQKvZq/?format=html&lang=pt. Acesso em: 27 abr. 2024.

RENDA média per capita no Brasil cresce 11,5% e atinge maior valor em 12 anos. Secretaria de Comunicação Social. Disponível em: https://www.gov.br/secom/pt-br/assuntos/noticias/2024/04/renda-media-per-capita-no-brasil-cresce-11-5-e-atinge-maior-valor-em-12-anos. Acesso em: 20 mar. 2025.

RODRÍGUEZ, Marcia Marina; BEZERRA, Byron Leite Dantas. Processamento de Linguagem Natural para Reconhecimento de Entidades Nomeadas em Textos Jurídicos de Atos Administrativos (Portarias). **Revista de Engenharia e Pesquisa Aplicada**, v. 5, n. 1, p. 67–77, 26 abr. 2020. ISSN 2525-4251. DOI: 10.25286/repa.v5i1.1204.

Disponível em:
http://revistas.poli.br/~anais/index.php/repa/article/view/1204. Acesso em: 11 mar. 2025.

SANTA ROSA, Lívia Neves; PIRES, Ana Clarissa dos Santos. **Leitelho: um coproduto versátil**. MilkPoint. 24 fev. 2021. Disponível em:
https://www.milkpoint.com.br/colunas/thermaufv/leitelho-um-coproduto-versatil-224170/.

SANTOS, Herlândia Cotrim. **Bebida fermentada sustentável "tipo iogurte grego" à base de leitelho**. 16 fev. 2023. Mestre em Ciência e Tecnologia de Alimentos – Universidade Federal de Viçosa, Viçosa - MG. DOI: 10.47328/ufvbbt.2023.256. Disponível em: https://locus.ufv.br//handle/123456789/31512. Acesso em: 21 mar. 2025.

SANTOS, Herlândia Cotrim *et al.* Enhancing dairy sustainability: Rheological, sensory, and physical-chemical properties of low-fat fermented beverages incorporating buttermilk. **Journal of Cleaner Production**, v. 443, p. 141159, mar. 2024. ISSN 0959-6526. DOI: 10.1016/j.jclepro.2024.141159. Disponível em:
https://www.sciencedirect.com/science/article/pii/S0959652624006061.

SERENGIL, Sefik Ilkin; OZPINAR, Alper. LightFace: A Hybrid Deep Face Recognition Framework. *In:* 2020 INNOVATIONS IN INTELLIGENT SYSTEMS AND APPLICATIONS CONFERENCE (ASYU). **2020 Innovations in Intelligent Systems and Applications Conference (ASYU)**. [*S. l.: s. n.*], out. 2020. p. 1–5. DOI: 10.1109/ASYU50717.2020.9259802. Disponível em:
https://ieeexplore.ieee.org/document/9259802. Acesso em: 3 abr. 2025.

SILEMG1935. **Dia Mundial do Queijo é celebrado dia 20 de janeiro**. Silemg. 21 jan. 2022. Disponível em: https://www.silemg.com.br/post/dia-mundial-do-queijo-e-celebrado-dia-20-de-janeiro. Acesso em: 10 jul. 2025.

SILVA, Femando Teixeira; CTAA, EMBRAPA. MANUAL DE PRODU • ÇA- O DE MANTEIGA, 1996.

SILVA, Tiago Aleff Da. UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC CURSO DE CIÊNCIA DA COMPUTAÇÃO, 2018.

SIQUEIRA, K. B. Um retrato do consumo de lácteos no Brasil., 2021. Accepted: 2021-09-13T01:01:41Z Publisher: Indústria de Laticínios, ano 25, n. 150, p. 58-59, 2021. Disponível em: http://www.infoteca.cnptia.embrapa.br/handle/doc/1134244. Acesso em: 25 mar. 2025.

SPITSBERG, V. L. Invited Review: Bovine Milk Fat Globule Membrane as a Potential Nutraceutical. English. **Journal of Dairy Science**, Elsevier, v. 88, n. 7, p. 2289–2294, 2005. ISSN 0022-0302. DOI: 10.3168/jds.S0022-0302(05)72906-4. Disponível em:

https://www.journalofdairyscience.org/article/S0022-0302(05)72906-4/fulltext.

STAILEY-YOUNG, Amos. **What's the Best Face Detector?** Python's Gurus. 28 jun. 2024. Disponível em: https://medium.com/pythons-gurus/what-is-the-best-face-detector-ab650d8c1225. Acesso em: 29 mar. 2025.

STEWART, Simon *et al.* **Selenium 2.0 Team**. Selenium. Section: documentation. Disponível em: https://www.selenium.dev/.

TAIGMAN, Yaniv *et al.* DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *In:* 2014 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **2014 IEEE Conference on Computer Vision and Pattern Recognition**. [*S. l.: s. n.*], jun. 2014. p. 1701–1708. ISSN: 1063-6919. DOI: 10.1109/CVPR.2014.220. Disponível em: https://ieeexplore.ieee.org/document/6909616. Acesso em: 5 jun. 2024.

TEIXEIRA, Isa Manuella Duarte *et al.* Elaboração de bebida à base de leitelho e análise sensorial de bebidas achocolatadas comerciais / Elaboration of buttermilk-based drink and sensory analysis of commercial chocolate drinks. **Brazilian Journal of Development**, v. 6, n. 6, p. 42010–42022, 30 jun. 2020. ISSN 2525-8761. DOI: 10.34117/bjdv6n6-658. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/12434. Acesso em: 27 abr. 2024.

TIROL. **Bebida Láctea Frutirol Coco 900g**. 2025. Disponível em: https://www.tirol.com.br/produto/bebida-lactea-frutirol-coco-900g/. Acesso em: 19 out. 2025.

VALOIS, Pedro Henrique Vaz. **Leveraging self-supervised learning for scene recognition in child sexual abuse imagery = Aprendizado auto-supervisionado para reconhecimento de cenas em imagens de abuso sexual infantil**. [*S. l.*], 2022. Disponível em: https://repositorio.unicamp.br/acervo/detalhe/1254557. Acesso em: 6 mar. 2025.

VAN ROSSUM, Guido; DRAKE, Fred L. **Python 3 Reference Manual**. Scotts Valley, CA: CreateSpace, fev. 2009. 242 p. ISBN 978-1-4414-1269-0.

VII ENAG AGROINDUSTRIA. **REAPROVEITAMENTO DE SUBPRODUTO LÁCTEO: ALTERNATIVAS PARA O USO DO LEITELHO**. [*S. l.: s. n.*], 16 dez. 2020. Disponível em: https://www.youtube.com/watch?v=G_5mctwFNow. Acesso em: 18 mar. 2024.

VIRTANEN, Pauli *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, v. 17, n. 3, p. 261–272, mar. 2020. Publisher: Nature

Publishing Group. ISSN 1548-7105. DOI: 10.1038/s41592-019-0686-2. Disponível em: https://www.nature.com/articles/s41592-019-0686-2. Acesso em: 30 out. 2024.

YANG, Lingling *et al.* Fusion of RetinaFace and improved FaceNet for individual cow identification in natural scenes. **Information Processing in Agriculture**, v. 11, n. 4, p. 512–523, dez. 2024. ISSN 22143173. DOI: 10.1016/j.inpa.2023.09.001. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S2214317323000653. Acesso em: 6 mar. 2025.