



# O que as redes sociais revelam sobre o consumo de lácteos

**Igor Infingardi de Carvalho Ribeiro**

JUIZ DE FORA

JANEIRO, 2026

# O que as redes sociais revelam sobre o consumo de lácteos

Igor Infingardi de Carvalho Ribeiro

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Wagner Antonio Arbex

JUIZ DE FORA

JANEIRO, 2026

# O que as redes sociais revelam sobre o consumo de lácteos

Igor Infingardi de Carvalho Ribeiro

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Wagner Antonio Arbex
Doutor em Engenharia de Sistemas e Computação

Luiz Maurilio da Silva Maciel
Doutor em Engenharia de Sistemas e Computação

Carlos Cristiano Hasenclever Borges
Doutor em Engenharia Civil

JUIZ DE FORA
20 DE JANEIRO, 2026

# Resumo

As redes sociais consolidaram-se como um importante canal para a expressão de opiniões sobre marcas, produtos e serviços, gerando grandes volumes de dados textuais não estruturados. No contexto da indústria de produtos lácteos, compreender essas manifestações é fundamental para identificar percepções, insatisfações e oportunidades de mercado.

Este trabalho tem como objetivo analisar comentários de consumidores sobre produtos lácteos publicados em redes sociais, combinando técnicas de análise de sentimentos e categorização temática. Para a análise de sentimentos, foi utilizado um classificador baseado em Support Vector Machine (SVM), treinado com comentários previamente rotulados. Já a categorização temática foi realizada por meio de chamadas à API da OpenAI, permitindo a identificação de diferentes temas presentes nos comentários, como preço, qualidade e modo de preparo.

Os resultados obtidos permitem uma análise integrada do sentimento associado aos comentários e os principais temas abordados pelos consumidores, fornecendo uma visão mais detalhada da percepção do público em relação aos produtos lácteos.

**Palavras-chave:** Análise de sentimentos, Redes Sociais, Lácteos, OpenAI, SVM.

# Abstract

Social media has consolidated itself as an important channel for the expression of opinions about brands, products, and services, generating large volumes of unstructured textual data. In the context of the dairy industry, understanding these manifestations is essential for identifying consumer perceptions, dissatisfactions, and market opportunities.

This study aims to analyze consumer comments on dairy products published on social media by combining sentiment analysis and thematic categorization techniques. For sentiment analysis, a classifier based on Support Vector Machine (SVM) was employed, trained on previously labeled comments. The thematic categorization was performed through calls to the OpenAI API, enabling multi-label classification of comments into different semantic categories, such as price, quality, and recipe preparation.

The results allow for an integrated analysis of comment polarity and the main topics discussed by consumers, providing a more detailed understanding of public perception regarding dairy products.



**Keywords:** Sentiment analysis, Social media, Dairy, OpenAI, SVM.

# Agradecimentos

Aos meus pais, Simone e Marcos Henrique, e ao meu irmão, Matheus, pelo apoio durante toda a graduação.

À minha namorada, Laura, por acreditar em mim e pelo apoio constante.

Ao meu orientador, Wagner Arbex, pela paciência, orientação e suporte ao longo do desenvolvimento deste trabalho.

Aos meus amigos da Computação, por tornarem a faculdade um ambiente mais leve, e aos professores do Departamento de Ciência da Computação, pelos ensinamentos transmitidos.

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviações

| | |
|---|---|
| DCC | Departamento de Ciência da Computação |
| UFJF | Universidade Federal de Juiz de Fora |
| NLP | Natural Language Processing |
| LLMs | Large Language Models |
| SVM | Support Vector Machine |
| API | Application Programming Interface |
| TF-IDF | Term Frequency - Inverse Document Frequency |

# 1 Introdução

## 1.1 Apresentação do Tema

Com o crescimento das redes sociais como ferramentas centrais de comunicação, interação e compartilhamento de experiências, tornou-se possível acessar um grande volume de dados gerados de forma espontânea por usuários sobre os mais diversos assuntos. No contexto do consumo, essas plataformas têm desempenhado um papel fundamental na formação da imagem pública de marcas e produtos, uma vez que os consumidores utilizam esses espaços para expressar opiniões, recomendar produtos e discutir suas experiências pessoais.

Entre os setores mais impactados por esse fenômeno está o da indústria alimentícia, em especial o segmento de produtos lácteos, que integra a rotina alimentar de grande parte da população. Itens como leite, iogurte e queijos são frequentemente mencionados em postagens que abordam temas como preço e qualidade.

Nesse cenário, compreender como os consumidores expressam suas opiniões sobre produtos lácteos nas redes sociais pode revelar os principais temas abordados e se o sentimento associado a ele é positivo ou negativo. Essas informações são valiosas para o desenvolvimento de novos produtos, aprimoramento de estratégias de marketing e identificação de oportunidades de mercado.

## 1.2 Contextualização

Com o avanço das técnicas de mineração de dados e do processamento de linguagem natural (NLP), tornou-se possível explorar grandes volumes de texto com o objetivo de extrair conhecimento útil para diferentes áreas. Entre essas técnicas, destacam-se a modelagem de temas e a análise de sentimentos, que vêm sendo aplicadas de forma crescente no ambiente digital, principalmente em redes sociais.

O uso dessas abordagens em contextos específicos, como o mercado de produtos

lácteos, ainda é pouco explorado, embora o setor apresente grande relevância econômica e social. Ao aplicar métodos computacionais ao conteúdo gerado por consumidores, é possível identificar tendências de consumo, temas em alta, e aceitação de produtos.

## 1.3 Motivação e Justificativa

As redes sociais transformaram-se em um importante espaço de interação entre consumidores e marcas, gerando diariamente um volume massivo de dados. De acordo com uma reportagem publicada pela revista *Forbes*, o Brasil é o terceiro maior consumidor de redes sociais em todo o mundo (PACETE, 2023), o que torna essas plataformas uma rica fonte de dados para estudos sobre comportamento do consumidor.

Como levantado no trabalho *Mineração de Dados em Rede Social para Avaliação de Tendências de Consumo do Queijo Artesanal no Brasil*, de Nogueira (NOGUEIRA, 2021), a ampla adoção das redes sociais pelos brasileiros tem incentivado empresas a acompanharem atentamente o que é publicado nesses ambientes, reconhecendo que as informações ali presentes são valiosas para a formulação de estratégias de mercado.

Diante disso, aplicar a modelagem de temas seguida da análise de sentimentos torna-se uma abordagem promissora para classificar e organizar os discursos presentes nas redes sociais. Essa sequência permite que empresas identifiquem os principais temas debatidos pelos consumidores, compreendam o sentimento geral associado a cada um deles e reajam de forma mais ágil às tendências e críticas do mercado.

## 1.4 Descrição do Problema

Apesar do crescente uso de redes sociais como fontes de dados para análise de comportamento do consumidor, ainda há uma lacuna quanto à aplicação sistemática de métodos de mineração de texto para compreender como os produtos lácteos são percebidos pelos usuários. A ausência de estudos aprofundados que combinem técnicas de modelagem de temas e análise de sentimentos nesse domínio específico limita o conhecimento sobre preferências, críticas e tendências do público.

O desafio consiste em identificar os principais temas discutidos pelos consumidores

e associar a eles sentimentos predominantes, com o objetivo de construir um panorama detalhado sobre a percepção social dos produtos lácteos no Brasil.

## 1.5 Questões de Pesquisa

A partir do problema descrito, estabelecem-se as seguintes questões norteadoras:

- Q1: Quais são os principais temas abordados pelos consumidores em postagens de redes sociais sobre produtos lácteos?
- Q2: Qual o sentimento predominante (positivo, negativo ou neutro) associado a cada um desses temas?

## 1.6 Objetivos

### 1.6.1 Objetivo Geral

Identificar a percepção dos consumidores em relação aos principais assuntos a respeito de produtos lácteos discutidos em redes sociais.

### 1.6.2 Objetivos Específicos

- Identificar os principais assuntos sobre lácteos presentes nas postagens em redes sociais.
- Identificar quais assuntos geram maior polarização emocional entre os usuários.
- Identificar o sentimento expresso nas postagens como positivo, negativo ou neutro.

# 2 Revisão Bibliográfica

A análise de sentimentos tem sido amplamente empregada como uma abordagem para compreender opiniões, percepções e emoções expressas por usuários em ambientes digitais. Com a consolidação das redes sociais como espaços de manifestação espontânea, diversos estudos passaram a explorar essas plataformas como fontes relevantes de dados, aplicando técnicas de processamento de linguagem natural em contextos sociais, políticos, econômicos e de consumo. No setor alimentício, em especial no contexto de produtos lácteos, essa abordagem tem se mostrado promissora para identificar percepções relacionadas à qualidade, confiança, saúde e preferências dos consumidores.

Esta revisão bibliográfica apresenta e discute trabalhos que aplicam técnicas de análise de sentimentos a dados extraídos de redes sociais e ambientes online, destacando os objetivos investigados, as fontes de dados utilizadas, os métodos empregados e as estratégias de rotulamento adotadas na literatura.

## 2.1 Objetivos dos estudos analisados

Diversos estudos têm utilizado a análise de sentimentos com o objetivo de compreender percepções públicas sobre temas de interesse coletivo. No contexto político e social, os trabalhos de SILVA; FARIA (2023) e YAGUI et al. (2017) exploraram dados do Twitter para analisar manifestações de opinião relacionadas a processos eleitorais e movimentos sociais, buscando identificar padrões de polarização e posicionamento ideológico.

Em aplicações voltadas ao monitoramento de marcas e serviços, EVANGELISTA; PADILHA (2014) analisaram postagens em redes sociais com o objetivo de avaliar a percepção de consumidores sobre empresas de e-commerce. De forma semelhante, o trabalho de OLENSCKI et al. (2020) investigou sentimentos associados ao uso da cloroquina, com foco na compreensão da reação pública frente a um tema amplamente debatido no campo da saúde.

No setor alimentício, o estudo PALEOLOGO et al. (2024) analisou percepções

sobre a qualidade do leite a partir de comentários publicados em redes sociais como Facebook e YouTube. O trabalho buscou identificar divergências entre a visão de consumidores, produtores e processadores, contribuindo para uma melhor compreensão das expectativas sociais relacionadas à cadeia produtiva do leite. Outros estudos, como os de HARBA; TIGU; DAVIDESCU (2021) e LINDQUIST et al. (2021), também exploraram sentimentos associados à alimentação, analisando emoções expressas em avaliações de restaurantes.

De forma geral, observa-se que a análise de sentimentos é empregada como uma ferramenta de apoio à tomada de decisão, permitindo compreender comportamentos sociais e identificar tendências a partir de dados gerados de forma espontânea por usuários.

## 2.2 Fontes de dados e redes sociais

As redes sociais configuram-se como a principal fonte de dados nos estudos analisados. O Twitter, atualmente denominado X, é amplamente utilizado, conforme observado nos trabalhos de SILVA; FARIA (2023), YAGUI et al. (2017), OLENSCKI et al. (2020) e SOUSA; FERNANDES (2023), sobretudo em razão da ampla popularidade da plataforma e da facilidade histórica de acesso aos dados publicados.

Entretanto, alterações recentes nas políticas de acesso à API da plataforma têm limitado a coleta de dados em larga escala, uma vez que consultas maiores passaram a exigir planos pagos. Esse cenário tem reduzido a viabilidade de estudos acadêmicos, especialmente aqueles conduzidos em contextos com restrições de recursos, o que tem levado pesquisadores a considerar fontes alternativas de dados em investigações mais recentes.

Embora o X seja a plataforma mais utilizada nos estudos analisados, outras redes sociais também têm sido amplamente empregadas. EVANGELISTA; PADILHA (2014) utilizaram dados do Facebook para monitorar percepções relacionadas ao comércio eletrônico, enquanto PALEOLOGO et al. (2024) combinaram dados do Facebook e do YouTube para analisar comentários sobre a qualidade do leite. Avaliações publicadas em plataformas especializadas, como o TripAdvisor, foram empregadas no estudo de HARBA; TIGU; DAVIDESCU (2021), possibilitando a análise de sentimentos no setor de restaurantes.

## 2.3 Técnicas, ferramentas e algoritmos

Sobre as técnicas de análise de sentimentos, observa-se uma diversidade de abordagens na literatura. Classificadores tradicionais de aprendizado de máquina, como Naive Bayes e Support Vector Machines (SVM), são amplamente utilizados, aparecendo em estudos como os de EVANGELISTA; PADILHA (2014), SOUSA; FERNANDES (2023) e PăVăLOIA et al. (2019). Esses métodos são frequentemente combinados com técnicas de pré-processamento textual, incluindo remoção de *stopwords*, lematização, normalização e tokenização.

Abordagens baseadas em léxicos também são recorrentes, com o uso de recursos como SentiWordNet e AFINN para a atribuição de polaridade aos textos. Essas técnicas são geralmente empregadas em análises exploratórias ou como complemento a modelos supervisionados.

Mais recentemente, observa-se a adoção crescente de modelos de linguagem pré-treinados. No estudo de PALEOLOGO et al. (2024), foi utilizado o modelo BERT para análise de sentimentos em comentários sobre leite, enquanto HENZ; HECKLER; BARBOSA (2025) empregaram o modelo Caramelo-Smile-2 em um contexto de saúde mental. Esses modelos têm se destacado por sua capacidade de capturar relações semânticas mais complexas, apresentando ganhos de desempenho quando comparados a abordagens tradicionais.

## 2.4 Estratégias de rotulamento dos dados

Os métodos de rotulamento utilizados nos estudos analisados variam conforme a disponibilidade de dados e os objetivos da pesquisa. O rotulamento manual é uma abordagem frequente, especialmente em estudos que analisam temas sensíveis ou contextos específicos, como eleições, movimentos sociais e debates sobre medicamentos, conforme observado nos trabalhos de SILVA; FARIA (2023), YAGUI et al. (2017) e OLENSCKI et al. (2020).

Outra estratégia recorrente consiste no uso de rótulos pré-existentes, extraídos de sistemas de avaliação baseados em estrelas, como no caso de plataformas de e-commerce e do TripAdvisor. Essa abordagem é explorada em estudos como os de EVANGELISTA;

PADILHA (2014) e HARBA; TIGU; DAVIDESCU (2021).

Por fim, alguns trabalhos utilizam modelos pré-treinados para realizar o rotulamento automático dos dados, como no estudo de PALEOLOGO et al. (2024), que emprega um modelo específico para análise de sentimentos em italiano. Essa estratégia tem se mostrado vantajosa em termos de escalabilidade e redução de esforço manual.

## 2.5 Síntese dos estudos revisados

A análise dos estudos revisados evidencia a consolidação da análise de sentimentos como uma abordagem relevante para a compreensão de opiniões expressas em ambientes digitais, sendo amplamente aplicada em diferentes contextos e domínios. De modo geral, observa-se a predominância do X como fonte de dados, contudo, outras plataformas também vêm sendo exploradas de acordo com as características e os objetivos de cada investigação.

Sobre as metodologias aplicadas, os trabalhos analisados indicam que técnicas tradicionais de aprendizado de máquina permanecem amplamente utilizadas. Entretanto, observa-se um crescimento no uso de modelos de linguagem pré-treinados, motivado pela capacidade desses modelos de capturar relações semânticas mais complexas e pela possibilidade de aplicação em bases de dados não rotuladas.

No contexto específico de produtos lácteos, os estudos ainda se mostram relativamente limitados. Esse cenário evidencia uma oportunidade para investigações adicionais que explorem, de forma mais aprofundada, as percepções dos consumidores a partir de dados provenientes de redes sociais, contribuindo para a ampliação do conhecimento na área.

# 3 Metodologia

Como ilustrado na Figura 3.1, a metodologia adotada neste trabalho é estruturada em um fluxo sequencial que abrange desde a disponibilização dos dados até a análise conjunta dos resultados. Na etapa inicial, foram utilizadas duas bases de dados previamente fornecidas, uma composta por comentários da rede social X e outra contendo comentários provenientes de vídeos do YouTube. Dessa forma, não foi necessária a realização de coleta direta de dados, apenas uma adaptação para juntar as duas bases.

Em seguida, os dados passam por uma etapa de limpeza e preparação, cujo objetivo é remover ruídos, padronizar os textos e garantir maior qualidade das informações analisadas. Após essa etapa, os comentários são submetidos a dois processos distintos e complementares, a análise de sentimentos e a categorização temática. A análise de sentimentos busca identificar a polaridade associada a cada comentário, enquanto a categorização temática permite identificar os principais temas abordados nos textos.

Por fim, os resultados dessas duas etapas são integrados em uma análise conjunta, possibilitando a avaliação do sentimento associado a cada categoria temática e permitindo uma visão mais detalhada da percepção dos consumidores sobre os produtos lácteos.

## 3.1 Coleta de dados

Para a realização deste trabalho, foram utilizadas duas bases de dados contendo comentários de redes sociais relacionados a produtos lácteos. Ambas as bases foram disponibilizadas pela Embrapa e não são de domínio público.

A primeira base é composta por comentários extraídos da plataforma X, contendo mais de 50 mil comentários relacionados a produtos lácteos. Cada comentário já está rotulado com o sentimento associado, sendo classificados como positivo (1), neutro (0) ou negativo (-1).

A segunda base é composta por comentários da plataforma YouTube, contendo mais de 30 mil comentários relacionados a produtos lácteos. Esta base foi definida como

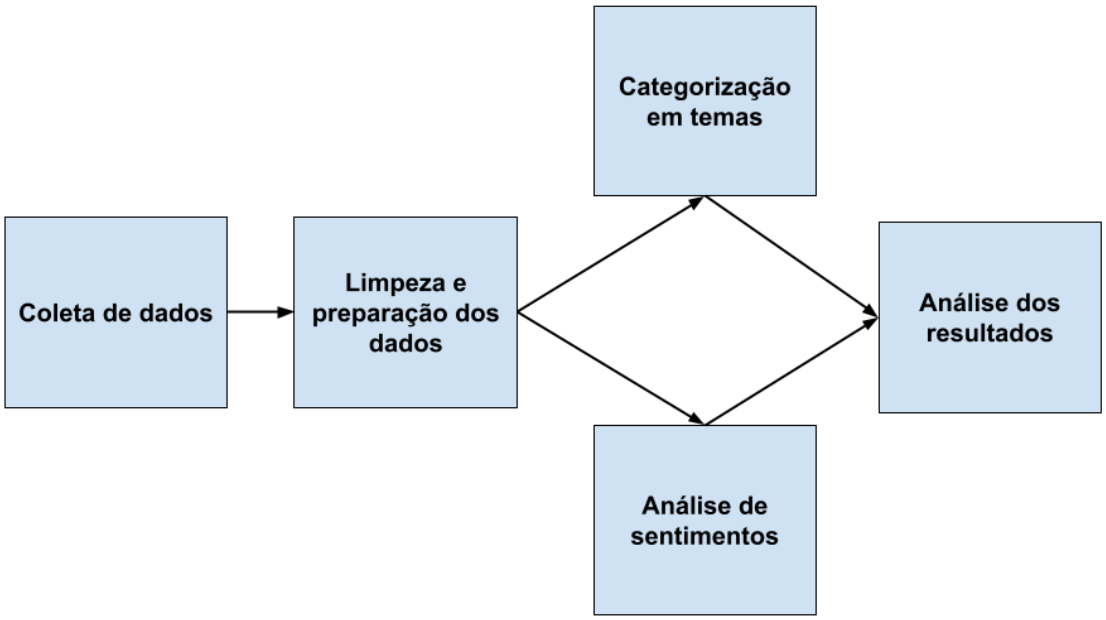


Figura 3.1: Fluxo de desenvolvimento do estudo.

a base-alvo deste trabalho, sendo utilizada tanto para a aplicação do modelo de análise de sentimentos quanto para o modelo de categorização temática dos comentários.

## 3.2 Limpeza e preparação dos dados

Antes da criação dos modelos, foi realizada uma etapa de limpeza e preparação dos dados, com o objetivo de garantir maior qualidade das informações analisadas. Para isso, foram aplicadas as seguintes restrições aos comentários:

- Remoção de comentários duplicados;
- Remoção de comentários contendo o caractere "?", uma vez que o foco do estudo está em opiniões expressas pelos usuários, e não em perguntas;
- Remoção de comentários com menos de cinco palavras, já que textos muito curtos dificilmente expressam opiniões completas ou informativas.

Para a etapa de análise de sentimentos, foram realizados pré-processamentos adicionais.

Inicialmente, foram removidas *stopwords*, que são palavras comuns, porém que não carregam muita informação, como “de“, “a“, “o“. Em seguida, foram removidos termos diretamente relacionados aos produtos analisados, como “leite”, “queijo” e “bolo”. Essa etapa teve como objetivo evitar que o modelo utilizasse o nome do produto como indício para a classificação do sentimento, forçando-o a basear sua decisão no conteúdo semântico do comentário.

## 3.3 Análise de sentimentos

Para a tarefa de análise de sentimentos, foram avaliados dois tipos de classificadores da biblioteca *scikit-learn*, amplamente utilizados em problemas desta natureza: o *LinearSVC*, baseado em *SVM*, e o *Multinomial Naive Bayes*, implementado por meio do classificador *MultinomialNB*.

Em ambos os cenários, os textos foram representados por meio da técnica de vetorização *Term Frequency - Inverse Document Frequency* (TF-IDF), que busca quantificar a relevância de cada termo em um documento em relação a toda a coleção de textos. Essa técnica combina a frequência de ocorrência de um termo em um documento com o inverso de sua frequência nos demais documentos, reduzindo a influência de palavras muito comuns e destacando termos mais discriminativos.

Além disso, foram selecionados e pré-processados 10 mil comentários para cada classe de sentimento (positivo, negativo e neutro). A divisão do conjunto de dados foi realizada seguindo a proporção de 70% para treinamento, 15% para validação e 15% para teste.

Embora o classificador SVM tenha sido originalmente proposto para problemas de classificação binária, sua aplicação em cenários multiclasse é viabilizada por meio de estratégias de decomposição do problema. No caso do modelo *LinearSVC*, a biblioteca *scikit-learn* adota, por padrão, a estratégia *One-vs-Rest*. Nessa abordagem, são treinados múltiplos classificadores binários independentes, cada um responsável por distinguir uma classe específica das demais. Considerando o problema abordado neste trabalho, com três

classes de sentimento, foram treinados três classificadores binários. Durante a etapa de predição, cada classificador gera uma pontuação associada à sua respectiva classe, sendo atribuída ao texto a classe correspondente à maior pontuação obtida.

O modelo *LinearSVC* foi configurado com o parâmetro $C = 1.0$, responsável por controlar o compromisso entre a maximização da margem e a penalização de erros de classificação. Adicionalmente, adotou-se a opção `dual=True`, apropriada para problemas em que o número de atributos é superior ao número de amostras, situação comum em representações textuais baseadas em TF-IDF.

O classificador *Multinomial Naive Bayes* foi empregado como modelo de comparação por ser amplamente utilizado em tarefas de classificação de textos. Esse método baseia-se no Teorema de Bayes e assume independência condicional entre os termos, considerando que a ocorrência de cada palavra em um documento é estatisticamente independente das demais, dado o rótulo da classe. Apesar dessa suposição simplificadora, o modelo apresenta bom desempenho em problemas de análise de sentimentos, especialmente quando aplicado a representações baseadas em frequência de termos.

O modelo *MultinomialNB* foi configurado com o parâmetro $\alpha = 1.0$, correspondente à suavização de Laplace. Essa técnica evita a atribuição de probabilidades nulas a termos ausentes em documentos de determinada classe durante o treinamento, contribuindo para maior estabilidade numérica e melhor capacidade de generalização do classificador. A escolha desse valor segue práticas amplamente adotadas na literatura.

Para a avaliação dos modelos, foram utilizadas três métricas amplamente empregadas em problemas de classificação: precisão, recall e F1-score. A precisão mede a proporção de predições corretas entre todas as instâncias classificadas como pertencentes a uma determinada classe, enquanto a revocação indica a capacidade do modelo de identificar corretamente todas as instâncias reais dessa classe. O *F1-score* corresponde à média harmônica entre precisão e revocação, fornecendo uma medida única e equilibrada do desempenho do classificador.

A Tabela 3.1 apresenta os resultados obtidos no conjunto de teste para ambos os classificadores.

Tabela 3.1: Desempenho dos classificadores SVM e Naive Bayes

| **Modelo** | **Classe** | **Precisão** | **Recall** | **F1-score** |
|---|---|---|---|---|
| SVM | Negativo | 0.87 | 0.84 | 0.85 |
| | Neutro | 0.80 | 0.88 | 0.84 |
| | Positivo | 0.88 | 0.82 | 0.85 |
| | Média | 0.85 | 0.85 | 0.85 |
| Naive Bayes | Negativo | 0.69 | 0.86 | 0.77 |
| | Neutro | 0.72 | 0.74 | 0.73 |
| | Positivo | 0.85 | 0.62 | 0.71 |
| | Média | 0.75 | 0.74 | 0.74 |

Observa-se que o classificador baseado em *SVM* apresenta desempenho superior na maioria das classes avaliadas. Por isso, ele foi selecionado para as análises subsequentes deste trabalho.

## 3.4 Categorização temática

Inicialmente, foi realizada uma análise manual exploratória dos primeiros 50 comentários da base de dados, com o objetivo de identificar os principais temas. A partir dessa análise inicial, foram definidas três categorias temáticas PRECO, QUALIDADE_GOSTO e SAUDE_NUTRICAO

Em uma segunda etapa, foi conduzido um processo de expansão do conjunto de categorias de forma manual, sem a utilização direta da API para classificação automática. Para isso, um subconjunto de comentários foi analisado com o apoio do *ChatGPT*, utilizado como ferramenta auxiliar para sugerir possíveis temas adicionais com base no conteúdo textual, resultando na adição das seguintes categorias INTOLERANCIA_LACTOSE, RECEITA_PREPARO, USO_CULINARIO, SUBSTITUICAO_PRODUTO, PROBLEMA_PRODUTO, DIFICULDADE_DUVIDA, PERCEPCAO_MARCA, OUTRO.

Após a definição final do conjunto de categorias, foi realizada a categorização automática dos comentários por meio de chamadas à API da *OpenAI*, utilizando o modelo *gpt-4.1-mini*. A escolha desse modelo deve-se ao fato dele equilibrar potência, desempenho

e preço acessível (GUINNESS, 2025).

A implementação da categorização automática foi realizada em Python, por meio de uma função responsável por enviar cada comentário de forma individual à API, fornecendo instruções ao modelo para atuar como um classificador semântico.

O processo de interação com o modelo foi estruturado por meio de dois tipos distintos de mensagens no prompt:

- system: Define regras gerais de funcionamento e orientações de comportamento do modelo, estabelecendo explicitamente que deveriam ser utilizadas somente as categorias fornecidas, que o número de categorias deveria estar entre uma e quatro e que o retorno deveria ser formatado em JSON

- user: Contém a mensagem que o modelo deve responder. Nela foram incluídos o comentário a ser analisado, bem como a lista completa de categorias e as descrições resumidas de cada uma delas, conforme apresentado na Tabela 3.2.

O uso dessa distinção de papéis segue práticas consolidadas em engenharia de prompt para modelos de linguagem, nas quais a mensagem de sistema é utilizada para configurar o contexto e orientar o comportamento do assistente, enquanto a mensagem de usuário representa a pergunta ou tarefa que o modelo deve interagir (HAKIM, 2025).

A definição de um número mínimo de uma categoria garantiu que todo comentário fosse classificado. A categoria *OUTRO* foi atribuída exclusivamente nos casos em que o comentário não apresentava correspondência semântica clara com nenhuma das demais categorias definidas, funcionando, portanto, como uma classe residual. Por sua vez, o limite máximo de quatro categorias teve como objetivo evitar a atribuição excessiva de rótulos, restringindo a classificação aos aspectos mais relevantes de cada comentário.

Devido ao tempo médio de resposta da API, estimado em aproximadamente 18 segundos por comentário, optou-se por limitar a análise a um subconjunto de 10 mil comentários da base do YouTube, garantindo viabilidade computacional sem comprometer a representatividade dos dados analisados.

Tabela 3.2: Categorias temáticas e descrições resumidas

| **Categoria** | **Descrição** |
|---|---|
| PRECO | Comentários sobre valor, custo, caro, barato. |
| QUALIDADE_GOSTO | Sabor, textura, resultado final. |
| SAUDE_NUTRICAO | Saúde, saudável, faz bem ou mal, engorda, conservantes, química, natural, industrializado, efeitos no corpo, composição alimentar. |
| INTOLERANCIA_LACTOSE | Alergia, sem lactose, desconforto. |
| RECEITA_PREPARO | Modo de fazer, ingredientes, processo. |
| USO_CULINARIO | Formas de consumo e aplicações. |
| SUBSTITUICAO_PRODUTO | Troca por outro produto. |
| PROBLEMA_PRODUTO | Estragou, talhou, defeitos. |
| DIFICULDADE_DUVIDA | Dúvidas, erros, dificuldades. |
| PERCEPCAO_MARCA | Confiança, credibilidade ou desconfiança. |
| OUTRO | Nenhum dos anteriores. |

As descrições apresentadas na Tabela 3.2 foram definidas de forma objetiva e resumida, sendo utilizadas como referência pelo modelo durante o processo de categorização. O objetivo dessas descrições é fornecer ao modelo uma noção semântico de cada categoria, sem aprofundar excessivamente seus significados, de modo a evitar direcionamentos ou vieses na classificação.

Em alguns casos específicos, como na categoria *SAUDE_NUTRICAO*, optou-se por uma descrição mais abrangente. Essa decisão foi tomada após observações indicarem que comentários relacionados a aspectos de saúde e nutrição não estavam sendo corretamente associados a essa categoria quando descrições mais curtas eram utilizadas.

# 4 Resultados

Este capítulo apresenta os resultados obtidos na base de comentários do YouTube. Primeiramente, é apresentada a distribuição geral dos sentimentos (Figura 4.1) e, em seguida, a distribuição das categorias temáticas identificadas (Figura 4.2). Por fim, os resultados são integrados por meio da análise de sentimentos por categoria, sintetizada no mapa de calor da Figura 4.3.

Os resultados analisados levam em conta somente os primeiros 10 mil comentários da base do YouTube que respeitam as regras definidas anteriormente, como não possuir o caractere "?" e não possuir comentários repetidos. A classificação de sentimentos foi realizada por meio do modelo baseado em *SVM*, treinado com a base previamente rotulada da plataforma X. Já a categorização temática foi obtida por meio de chamadas à API da *OpenAI*, utilizando o modelo *gpt-4.1-mini* para atribuição multirrótulo das categorias definidas na metodologia.

## 4.1 Análise de sentimentos

Como ilustrado na Figura 4.1, observa-se que aproximadamente 60,3% dos comentários analisados foram classificados como positivos, indicando uma percepção majoritariamente favorável dos consumidores em relação aos produtos lácteos no contexto do YouTube.

## 4.2 Categorização em temas

A Figura 4.2 apresenta a distribuição das categorias temáticas identificadas nos comentários. Observa-se predominância das categorias relacionadas ao preparo de receitas e à qualidade dos produtos, o que é esperado considerando o contexto da plataforma YouTube, amplamente utilizada para compartilhamento de conteúdos culinários.

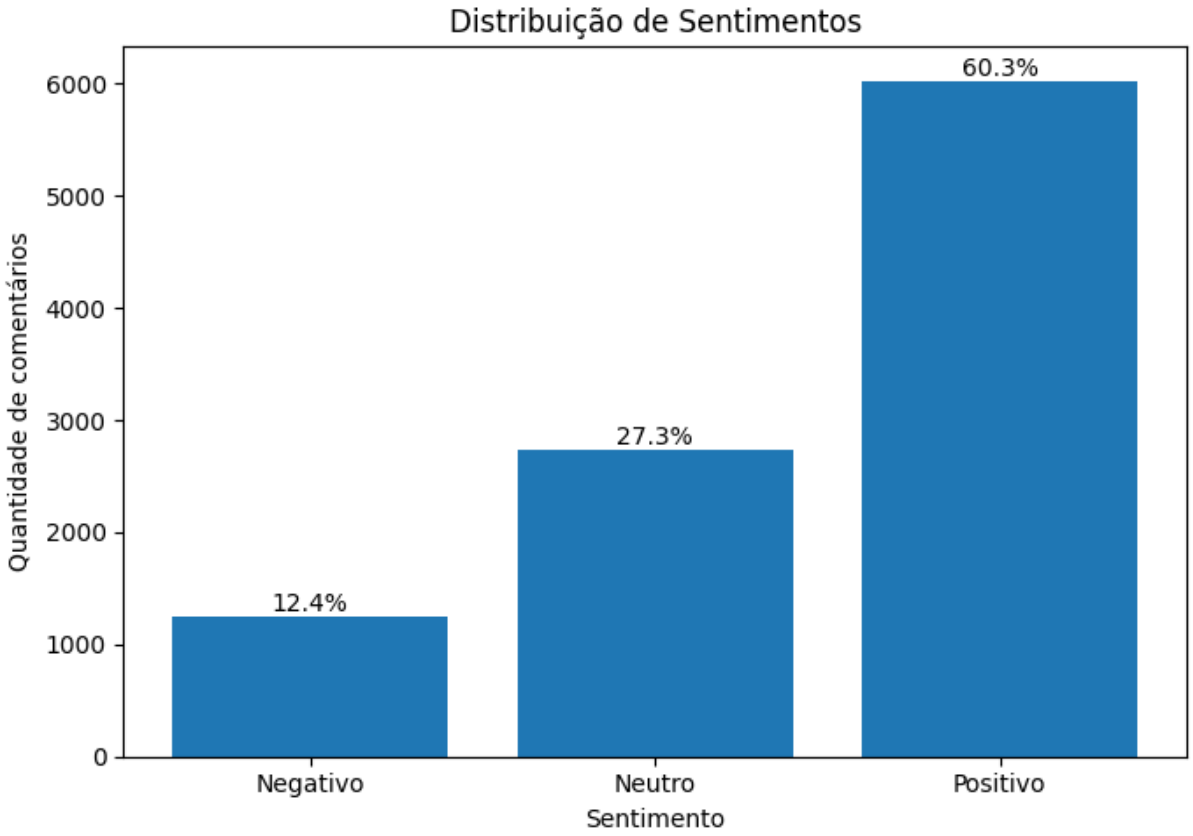


Figura 4.1: Distribuição de sentimentos.

## 4.3 Resultados consolidados

Esses dois modelos nos ajudam a trazer insights importantes sobre os comentários de produtos lácteos, porém juntos eles podem trazer informações muito mais valiosas, como por exemplo uma análise de sentimentos por categoria.

O mapa de calor apresentado na Figura 4.3 consolida os resultados da análise de sentimentos por categoria temática. Observa-se que a categoria QUALIDADE_GOSTO apresenta a maior proporção de comentários positivos, enquanto a categoria PROBLEMA_PRODUTO concentra o maior percentual de sentimentos negativos.

Além disso, a categoria *PREÇO* apresenta uma parcela significativa de comentários negativos, indicando uma possível insatisfação dos consumidores em relação ao custo dos produtos. Esses resultados evidenciam como a análise conjunta de sentimentos e categorias temáticas pode fornecer insights relevantes para empresas interessadas em compreender a percepção do público e identificar oportunidades de melhoria ou estratégias de mercado.

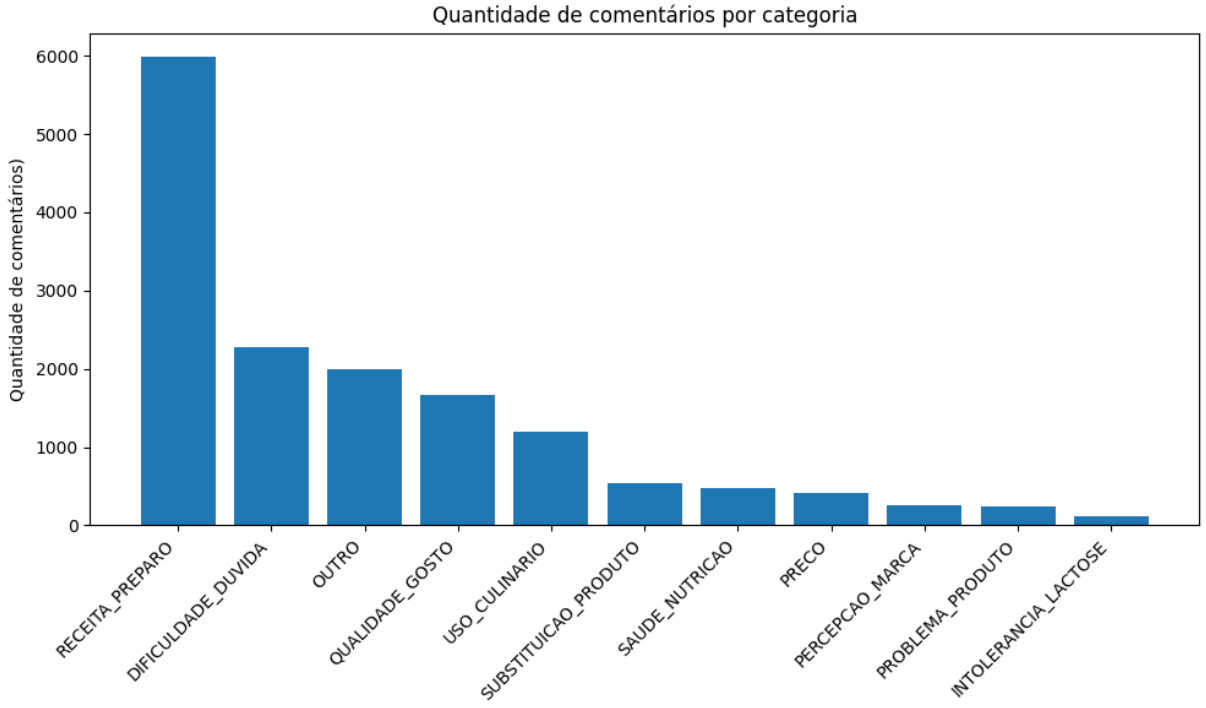


Figura 4.2: Distribuição de categorias.

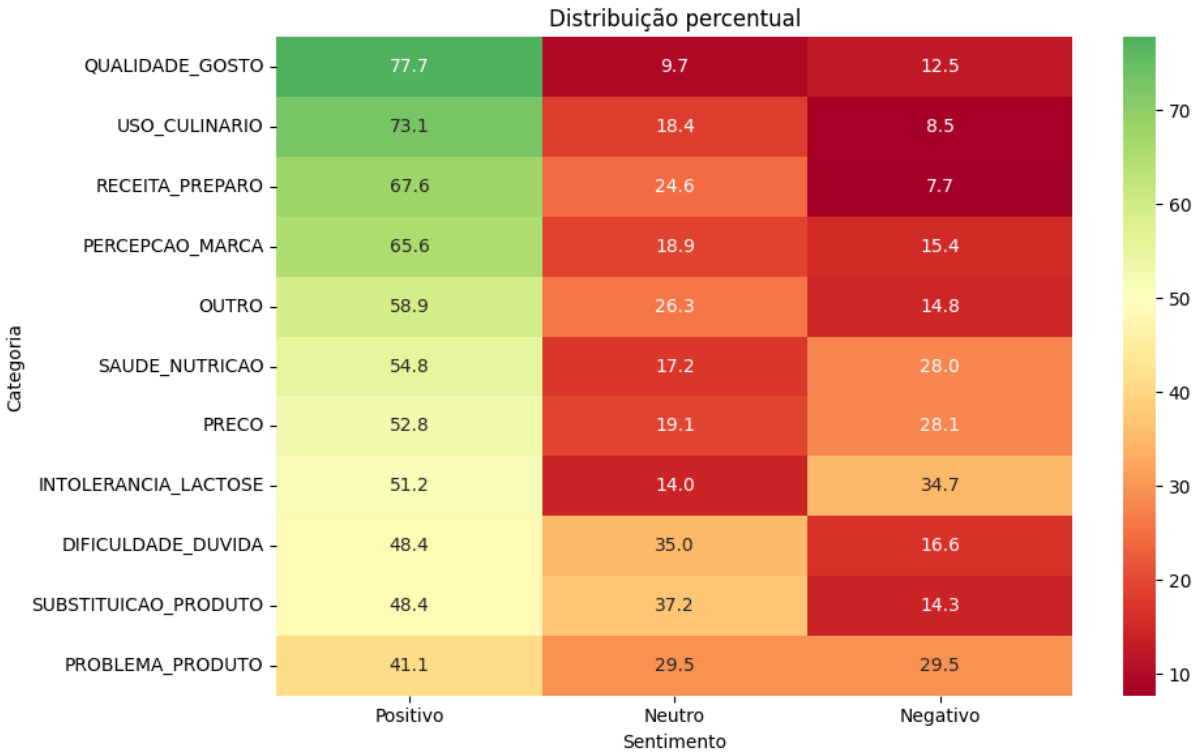


Figura 4.3: Heatmap da análise de sentimento por categoria.

# 5 Conclusão

Este trabalho teve como objetivo geral investigar a percepção dos consumidores a respeito de produtos lácteos a partir de comentários extraídos de redes sociais, por meio da integração de técnicas de análise de sentimentos e categorização temática baseadas em Processamento de Linguagem Natural. A partir dos resultados obtidos, observa-se que esse objetivo foi plenamente atingido, uma vez que foi possível identificar os principais temas discutidos pelos consumidores e analisar a polaridade emocional associada a eles.

Em relação aos objetivos específicos, o primeiro consistiu em identificar e classificar automaticamente os principais assuntos abordados nos comentários sobre produtos lácteos. Esse objetivo foi alcançado por meio do processo de categorização temática, que evidenciou como temas mais recorrentes o preparo de receitas, dificuldades e dúvidas durante o uso, qualidade e sabor do produto, uso culinário e substituição de ingredientes. Esses resultados são coerentes com o contexto da plataforma analisada, uma vez que a maior parte dos comentários do YouTube está associada a vídeos de preparo de receitas, nos quais os usuários compartilham experiências práticas, dúvidas e sugestões relacionadas ao uso dos produtos.

O segundo objetivo específico, voltado à identificação do sentimento expresso nas postagens, também foi atendido com sucesso. A aplicação do classificador *SVM*, treinado a partir de uma base previamente rotulada da plataforma X, permitiu classificar os comentários do YouTube quanto à polaridade emocional. Os resultados indicaram que 60,3% dos comentários apresentam sentimento positivo, 27,3% são neutros e apenas 12,4% expressam sentimento negativo, evidenciando uma percepção majoritariamente favorável dos consumidores em relação aos produtos lácteos analisados.

Por fim, o terceiro objetivo específico buscou identificar quais categorias temáticas apresentam maior polarização emocional. A análise integrada entre categorização temática e sentimentos revelou que algumas categorias concentram percentuais elevados de sentimentos negativos, destacando-se intolerância a lactose, com 34,7% de comentários negativos, problema relacionado ao produto, com 29,5%, e preço, com 28,1%. Por outro

lado, categorias como qualidade gosto, uso culinário e preparo de receitas apresentaram predominância de sentimentos positivos, com 77,7%, 73,1% e 67,6% dos comentários, respectivamente. Esses resultados indicam que, embora a percepção geral dos consumidores seja positiva, existem aspectos específicos que geram maior insatisfação, evidenciando oportunidades de melhoria, especialmente no desenvolvimento de produtos voltados a consumidores com intolerância à lactose, na redução de problemas relacionados à qualidade e na adequação dos preços.

De modo geral, os resultados obtidos demonstram que a combinação entre modelos clássicos de aprendizado de máquina, como o *SVM*, e modelos de linguagem de grande porte constitui uma abordagem eficaz para a análise de opiniões em redes sociais. Essa integração permite não apenas identificar o sentimento global dos consumidores, mas também compreender os diferentes aspectos que influenciam suas percepções, fornecendo subsídios relevantes para empresas e instituições interessadas em monitorar a opinião pública e identificar oportunidades de aprimoramento.

Apesar dos resultados promissores, algumas limitações devem ser consideradas. O uso da API da OpenAI impôs restrições quanto ao volume de dados analisados, em função do custo computacional e do tempo de resposta. Além disso, a categorização temática depende da interpretação semântica do modelo de linguagem, o que pode introduzir variações difíceis de controlar integralmente. Ainda assim, o estudo evidencia o potencial da abordagem proposta e reforça a viabilidade do uso integrado de técnicas de PLN para a análise de grandes volumes de dados textuais não estruturados no contexto de produtos lácteos.

## 5.1 Trabalhos Futuros

Como trabalhos futuros, diversas extensões podem ser exploradas a partir dos resultados obtidos neste estudo. Uma possibilidade consiste em realizar uma análise mais aprofundada sobre a viabilidade de utilizar o classificador de análise de sentimentos treinado na base da plataforma X diretamente na base do YouTube. Essa investigação pode incluir avaliações comparativas de desempenho, bem como a adaptação ou o reentreinamento do modelo considerando as particularidades linguísticas e contextuais de cada plataforma.

Outra direção relevante envolve o treinamento de modelos de análise de sentimentos específicos para cada categoria temática identificada, uma vez que um mesmo comentário pode expressar opiniões distintas sobre diferentes aspectos do produto, como avaliações positivas relacionadas à qualidade e negativas associadas ao preço. Essa abordagem permitiria uma análise mais refinada e precisa da percepção dos consumidores.

Adicionalmente, propõe-se a realização de análises comparativas entre diferentes empresas ou marcas do setor lácteo. Essa extensão possibilitaria identificar variações na percepção dos consumidores entre concorrentes, bem como mapear pontos fortes e fragilidades específicas de cada empresa em relação a categorias como qualidade, preço, problemas no produto e intolerância à lactose, fornecendo subsídios relevantes para estratégias de posicionamento e tomada de decisão.

Por fim, destaca-se como trabalho futuro a validação sistemática do uso de modelos de linguagem de grande porte, como o *GPT*, enquanto classificadores automáticos de categorias temáticas. Essa validação pode ser conduzida por meio da comparação entre as classificações atribuídas pelo modelo e anotações manuais realizadas por avaliadores humanos, utilizando métricas como acurácia e medidas de concordância. Tal análise permitiria avaliar de forma objetiva a confiabilidade, a consistência e eventuais vieses do modelo no processo de categorização temática.

# Bibliografia

EVANGELISTA, T. R.; PADILHA, T. P. P. Monitoramento de posts sobre empresas de e-commerce em redes sociais utilizando análise de sentimentos. In: *Anais do 3º Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*. Brasília: Sociedade Brasileira de Computação, 2014. p. 152–163. ISSN 2595-6094.

GUINNESS, H. *Modelos OpenAI: Cada modelo e para que ele é mais indicado.* 2025. Disponível em: ⟨https://zapier.com/blog/openai-models/⟩.

HAKIM, M. *Mastering Prompt Engineering: A Guide to System, User, and Assistant Roles in OpenAI API*. 2025. Disponível em: ⟨https://medium.com/@mudassar.hakim/mastering-prompt-engineering-a-guide-to-system-user-and-assistant-roles-in-openai-api-28fe5fbf1d8

HARBA, J.-N.; TIGU, G.; DAVIDESCU, A. A. Exploring consumer emotions in pre-pandemic and pandemic times. a sentiment analysis of perceptions in the fine-dining restaurant industry in bucharest, romania. *International Journal of Environmental Research and Public Health*, v. 18, n. 24, p. 13300, 2021. Disponível em: ⟨https://doi.org/10.3390/ijerph182413300⟩.

HENZ, M. D.; HECKLER, W. F.; BARBOSA, J. L. V. Uma avaliação da capacidade de modelos de linguagem para análise de sentimentos em um contexto de saúde mental. In: *Anais do 25º Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*. Porto Alegre/RS: Sociedade Brasileira de Computação, 2025. p. 293–304. ISSN 2763-8952. Disponível em: ⟨https://doi.org/10.5753/sbcas.2025.7075⟩.

LINDQUIST, J. et al. Food for thought: A natural language processing analysis of the 2020 dietary guidelines public comments. *The American Journal of Clinical Nutrition*, v. 114, n. 2, p. 713–720, 2021. Disponível em: ⟨https://doi.org/10.1093/ajcn/nqab119⟩.

NOGUEIRA, T. S. *Mineração de Dados em Rede Social para Avaliação de Tendências de Consumo do Queijo Artesanal no Brasil*. Dissertação (Dissertação de Mestrado) — Universidade Federal de Juiz de Fora, Juiz de Fora, 2021.

OLENSCKI, J. et al. Aplicação de análise de sentimentos no twitter para avaliação da percepção pública quanto a cloroquina. In: *Anais do 20º Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*. Evento Online: Sociedade Brasileira de Computação, 2020. p. 500–505. ISSN 2763-8952. Disponível em: ⟨https://doi.org/10.5753/sbcas.2020.11547⟩.

PACETE, L. G. *Brasil é o terceiro país que mais consome redes sociais em todo o mundo*. 2023. *Forbes Brasil*. Disponível em: ⟨https://forbes.com.br/forbes-tech/2023/03/brasil-e-o-terceiro-pais-que-mais-consome-redes-sociais-em-todo-o-mundo/⟩. Acesso em: 18 jul. 2025.

PALEOLOGO, M. et al. Exploring social media to understand perceptions of milk quality among farmers, processors, and citizen-consumers. *Foods*, v. 13, n. 16, p. 2526, 2024. Disponível em: ⟨https://doi.org/10.3390/foods13162526⟩.

PăVăLOIA, V.-D. et al. Opinion mining on social media data: Sentiment analysis of user preferences. *Sustainability*, v. 11, n. 16, p. 4459, 2019. Disponível em: ⟨https://doi.org/10.3390/su11164459⟩.

SILVA, S. M. B.; FARIA, E. R. Análise de sentimentos expressos no twitter em relação aos candidatos da eleição presidencial de 2022. In: *Anais do 12º Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*. João Pessoa/PB: Sociedade Brasileira de Computação, 2023. p. 79–90. ISSN 2595-6094. Disponível em: ⟨https://doi.org/10.5753/brasnam.2023.229992⟩.

SOUSA, T. M. de; FERNANDES, D. S. A. Expansão automática de léxico para análise de sentimentos de twitter no domínio do mercado financeiro brasileiro. In: *Anais da 11ª Escola Regional de Informática de Goiás (ERI-GO)*. Goiânia/GO: Sociedade Brasileira de Computação, 2023. Disponível em: ⟨https://doi.org/10.5753/erigo.2023.237321⟩.

YAGUI, M. M. M. et al. "bela, recatada e do lar": Base de dados e aspectos do movimento social ocorrido na rede social online twitter. In: *Anais do 14º Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*. São Paulo: Sociedade Brasileira de Computação, 2017. p. 1585–1594. ISSN 2326-2842. Disponível em: ⟨https://doi.org/10.5753/sbsc.2017.9951⟩.