JCEC
THE JOURNAL OF ENGINEERING AND EXACT SCIENCES

OPEN ACCESS

The Journal of Engineering and Exact Sciences – jCEC, Vol. 08 N. 10 (2022)
journal homepage: https://periodicos.ufv.br/jcec
eISSN: 2527-1075
ISSN: 2446-9416

Data mining in tweets for analyzing dairy consumption in Brazil

Mineração de dados em *tweets* para análise do consumo de lácteos no Brasil

Article Info:
Article history: Received 2022-11-10 / Accepted 2022-12-01 / Available online 2022-12-01
doi: 10.18540/jcecvl8iss10pp14863-01a

Thallys da Silva Nogueira
ORCID: https://orcid.org/0000-0002-8499-0181
Universidade Federal de Juiz de Fora, Brasil
E-mail: thallys.nogueira@estudante.ufjf.br
Anna Letícia Franco Monteiro
ORCID: https://orcid.org/0000-0002-0038-349X
Universidade Federal de Juiz de Fora, Brasil
E-mail: anna.franco@estudante.ufjf.br
Darlan Henrique da Costa Silva
ORCID: https://orcid.org/0000-0001-5284-3972
Universidade Federal de Juiz de Fora, Brasil
E-mail: darlan.silva@ice.ufjf.br
Kennya Beatriz Siqueira
ORCID: https://orcid.org/0000-0001-6727-7774
Empresa Brasileira de Pesquisa Agropecuária, Brasil
E-mail: kennya.siqueira@embrapa.br
Priscila Vanessa Zabala Capriles Goliatt
ORCID: https://orcid.org/0000-0001-9780-4328
Universidade Federal de Juiz de Fora, Brasil
E-mail: capriles@ice.ufjf.br

Resumo
A pandemia da COVID-19 causou diversos impactos na rotina dos brasileiros e a alimentação é um deles. O foco deste trabalho foi analisar o consumo de produtos lácteos no Brasil nos últimos tempos, empregando dados da rede social Twitter, com a ferramenta Observatório do Consumidor. Com o objetivo de responder às perguntas "Quais são os derivados lácteos mais consumidos no Brasil?" e "Como foi este consumo ao longo do tempo?", utilizou-se técnicas de processamento de linguagem natural nos dados para identificar os verbos referentes ao consumo e suas respectivas frequências ao longo do tempo. Foi observado que sorvete, leite condensado, queijos, doce de leite e leite foram os cinco produtos lácteos que obtiveram maior número de menções a verbos que remetem ao consumo, caracterizando-os como os produtos mais consumidos no período analisado. Entretanto, foi observado que o consumo de lácteos vem diminuindo desde 2020. Estes resultados mostram que é possível analisar de forma rápida, dinâmica e barata, o consumo de alimentos por meio das redes sociais.
Palavras-chave: Consumidor. Leite e derivados. Inteligência artificial. Redes sociais. Pesquisa de mercado.

Abstract
Brazilians' daily routine was affected by the COVID-19 epidemic in a number of ways, with food being one of them. This study used data from the social network Twitter and the tool Observatório do Consumidor to examine the consumption of dairy products in Brazil in recent years. Natural

language processing techniques were applied to the data to determine the verbs relating to consumption and their respective frequencies over time in order to respond to the queries "Which are the most consumed dairy products in Brazil?" and "How was this consumption over time?". It was found that the five dairy products with the largest number of consumption-related verb references were ice cream, condensed milk, cheese, dulce de leche, and milk, making them the most popular choices during the study period. However, it was noted that since 2020, dairy consumption has been declining. These findings demonstrate that it is feasible to swiftly, dynamically, and affordably assess food consumption through social networks.
Keywords: Consumer. Milk and derivatives. Artificial intelligence. Social networks. Market research.

1. Introdução

Coletar, armazenar, processar e analisar dados provenientes das redes sociais sobre o consumo dos derivados lácteos é o que o Observatório do Consumidor (OC) vem realizando desde o ano de 2020. Desenvolvido através da parceria entre as instituições Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA Gado de Leite, Universidade Federal de Juiz de Fora (UFJF) e o Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas - Campus Juiz de Fora (IF-Sudeste/MG), o OC é uma plataforma que busca desenvolver alternativas às pesquisas de mercado tradicionais. Por meio do uso de técnicas de inteligência artificial, mineração de dados, *web*-semântica e processamento de linguagem natural, essa ferramenta identifica informações sobre o perfil dos consumidores de lácteos no Brasil e as principais características de consumo de forma rápida, barata e representativa sendo estes um dos principais pontos de vantagem no uso do OC para confecção de pesquisas de mercado (Nogueira, 2021).

Desde maio de 2020, semanalmente, em média 165.617 publicações do Twitter, denominadas *tweets*, são coletadas referente a 10 categorias de lácteos[1]. Desde a criação do OC, diversos trabalhos foram realizados, explorando de diferentes maneiras o conteúdo publicado sobre os lácteos no Twitter. O primeiro caso de estudo do OC foi relacionado à mudança no comportamento de consumo de derivados lácteos no Brasil antes e durante a pandemia por COVID-19 (Siqueira *et al.*, 2020a; Siqueira *et al.*, 2020b). Em outro momento, NOGUEIRA *et. al.* (2022) (Nogueira *et al.*, 2022) analisam como foi o consumo de queijo artesanal no Brasil.

Neste sentido, o foco deste trabalho está relacionado com a mineração de informações contidas em dados coletados do Twitter sobre os lácteos no Brasil. Com auxílio do OC, o objetivo do trabalho foi compreender o consumo de lácteos no Brasil identificando os mais consumidos no período compreendido entre 07/05/2020[2] e 11/08/2022.

2. Material e Métodos

A arquitetura do OC é organizada em 3 principais módulos que são (i) coleta e armazenamento de dados, (ii) processamento de dados e por fim (iii) mineração de dados e extração de informações. O fluxo de tarefas (1) é iniciado com a seleção das palavras-chave (nome dos produtos lácteos de interesse) e submissão das mesmas ao algoritmo de coleta e armazenamento que grava cada uma das publicações em um banco de dados relacional desenvolvido. Para extrair informações relevantes destes dados, utiliza-se técnicas de processamento de linguagem natural em que realiza-se a padronização de cada um dos *tweets*. Nesta etapa, realiza-se a remoção de *stopwords*, *emojis* e *links* de *sites* que de forma recorrente aparecem nos *tweets* e não carregam

[1] Bebidas Lácteas, Creme de Leite, Doce de Leite, Iogurte, Leite, Leite Condensado, Leite Fermentado, Manteiga, Queijos e Sorvete.

[2] Início da coleta de dados do Observatório do Consumidor.

informação interessante consigo. Com os dados padronizados, inicia-se então a mineração dos dados para a extração das informações de interesse.

Utilizou-se a linguagem de programação Python (Van Rossum *et al.*, 2009), em conjunto com diversas bibliotecas sendo a *spaCy* (Honnibal *et al.*, 2017), NLTK (*Natural Language Toolkit*) (Bird *et al.* 2009) as mais utilizadas. Ambas foram empregadas para extração de informações referentes ao consumo dos lácteos no conteúdo dos *tweets*.

Utilizando a biblioteca *spaCy*, mapeou-se a ocorrência da relação sintática entre "verbo" e "substantivo" em todos os *tweets* coletados com o objetivo de identificar os principais verbos e substantivos que indicam respectivamente o ato de consumir e os nomes dos lácteos que são analisados pelo OC.

3. Resultados e Discussão

Ao todo 19.708.498 *tweets* sobre lácteos foram coletados pelo OC. Destes, 70,4% dos *tweets* apresentaram pelo menos uma ação no contexto dos lácteos e possuíam ao menos um verbo, podendo este ser referente ao consumo ou não. Como o objetivo deste trabalho é o de extrair informações sobre o consumo, analisou-se em específico a ocorrência de cinco principais verbos que remetem ao consumo com suas respectivas flexões verbais: comer, beber, tomar, comprar e fazer. Para garantir que de fato houve o consumo de lácteos foram considerados verbos flexionados nos tempos verbais do passado e do presente.

Após realizar a filtragem dos dados, 2.410.272 *tweets* foram selecionados por apresentarem algum dos verbos de interesse. Os resultados mostraram que os cinco lácteos mais consumidos foram: sorvete (com 851.917 *tweets*), leite condensado (541.972 *tweets*), queijos (257.311 *tweets*), doce de leite (227.568 *tweets*) e leite (101.176 *tweets*). A Figura 1 mostra, em detalhes, a proporção de menções a cada um dos verbos de consumo por lácteo.

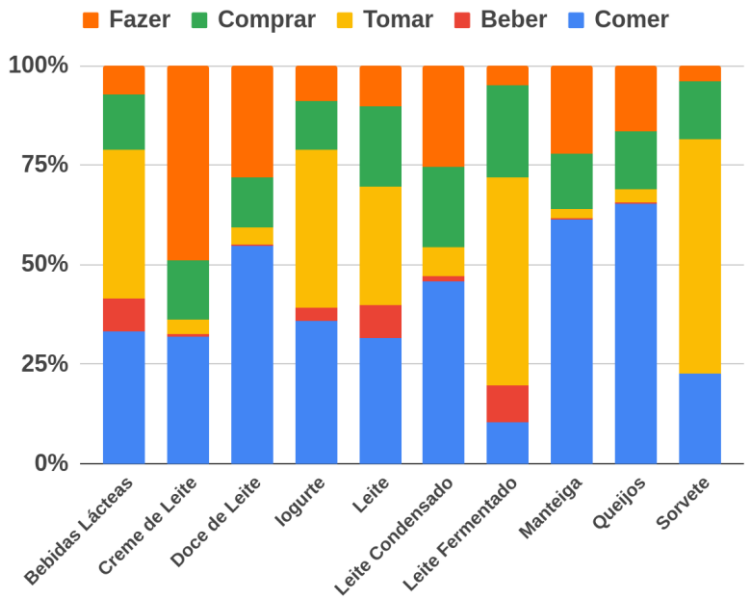

Figura 1 - Proporção de menções aos verbos de consumo por derivado lácteo no período de 07/05/2020 e 11/08/2022. Retirado da rede social Twitter. Fonte: resultados desta pesquisa.

Cada tipo de produto possui seu próprio modo de consumo. Da Figura 1 é possível verificar que para o verbo "fazer", que está diretamente relacionado ao ato de seguir os passos requeridos em receitas culinárias, os produtos creme de leite, doce de leite e leite condensado foram os que obtiveram maior representatividade com 67,2% de todas menções a este verbo. Para a ação verbal

"comprar", os produtos sorvete, leite condensado e queijos foram os 3 produtos com mais menções, com 35%, 31,3% e 10,7% respectivamente. Para o verbo "tomar", os produtos sorvete, leite fermentado, iogurte, bebidas lácteas e leite tiveram os maiores números de menções a essa ação. Vale destacar que o sorvete foi consideravelmente o produto mais "tomado" no tempo em questão, com 81% dessa ação. O verbo "beber" não obteve uma expressão tão significativa como os demais, entretanto, os produtos lácteos líquidos, ou seja, leite fermentado, bebidas lácteas, leite e iogurte, foram os mais expressivos em número de menções a esse verbo. Por fim, o verbo "comer" esteve presente em todos os produtos analisados, sendo os de maior destaque: queijos, manteiga, doce de leite e leite condensado. As análises dos *tweets* mostraram que os queijos foram mais consumidos isoladamente ou em lanches *delivery e* a manteiga esteve presente em receitas e no café da manhã. Já o doce de leite e o leite condensado, produtos de caráter indulgente, foram amplamente consumidos em diversos momentos e opções de consumo, mostrando como esses lácteos fizeram parte da rotina alimentar dos brasileiros durante o período avaliado.

Os resultados alcançados pela análise destes dados corroboram e confirmam os que foram relatados em ((Siqueira *et al.*, 2020a; Siqueira *et al.*, 2020b). Os produtos lácteos mais consumidos são caracterizados indulgentes e o pico de consumo destes produtos ocorreu no período compreendido entre maio e dezembro de 2020 como pode ser observado na Figura 2.

Da Figura 2, é possível verificar que o consumo de lácteos foi decaindo ao longo dos meses, fato este que tem sido confirmado atualmente pelo setor (Ruggiero, 2022). Dos 5 lácteos em destaque, o sorvete foi o produto mais consumido com um pico de consumo em outubro de 2020 quando alcançou 58.357 menções aos verbos de consumo.

O leite condensado apresentou um pico em janeiro de 2021. No entanto, esse *outlier* foi ocasionado por um fato bastante comentado nas redes sociais, referente à divulgação do elevado valor de compras do produto feito pelo governo federal.

O produto leite apresentava um comportamento bastante estável ao longo dos meses. Mas, em setembro de 2021, começa a obter maior destaque. Isso ocorreu devido à adição de novas palavras-chave referentes a produtos que até então não eram explorados pelo OC.

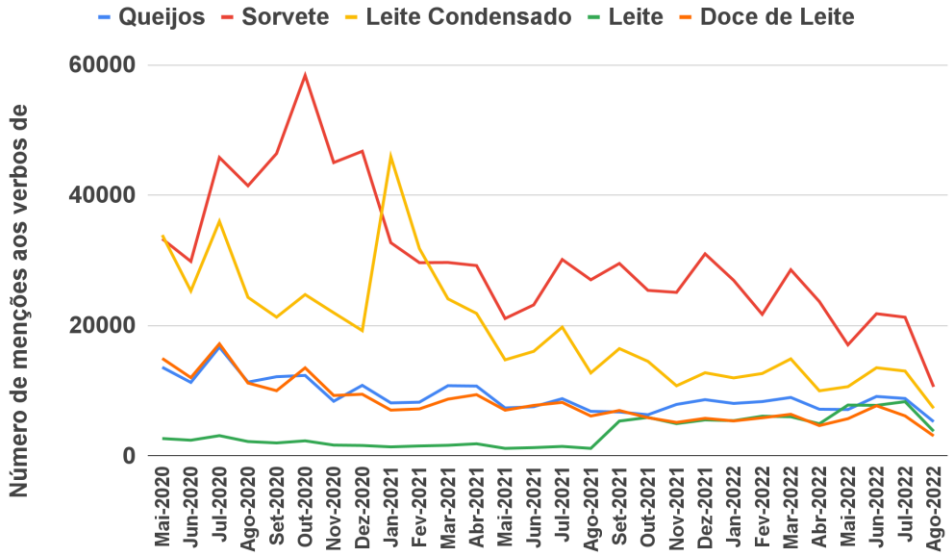

Figura 2 - Número de menções aos verbos de consumo ao longo do tempo dos 5 derivados lácteos mais representativos. Retirado da rede social Twitter. Fonte: resultados desta pesquisa.

4. Conclusões

Esse trabalho mostrou que a coleta, o processamento e a análise de dados de rede social feita pelo Observatório do Consumidor inova a forma de responder a perguntas de uma pesquisa de mercado. Atuando nas deficiências da pesquisa de mercado tradicional, o OC possibilita a realização de análises mais representativas em uma grande quantidade de dados por todo o território nacional sem a necessidade da aplicação de questionários. Com essa ferramenta foi possível identificar que os derivados lácteos mais consumidos no Brasil foram: sorvete, leite condensado, queijos, doce de leite e leite. Além disso, a pesquisa mostrou que o consumo de lácteos vem decaindo desde 2020. Para trabalhos futuros espera-se ampliar as questões a serem respondidas, além de explorar dados de outras redes sociais.

Referências

Nogueira, T. S. (2021). Mineração de dados em rede social para avaliação de tendências de consumo do queijo artesanal no Brasil. Dissertação de Mestrado. Universidade Federal de Juiz de Fora, MG, Brasil.

Siqueira, K. B. *et al.* (2020). Análise exploratória da imagem dos lácteos em tempos de coronavírus. Indústria de laticínios, n. 143, p. 64-66, ISSN 1678-7250.

Siqueira, K. B. *et al.* (2020). O impacto da pandemia no consumo de lácteos no Brasil. Indústria de laticínios, n. 147, p. 36-38. ISSN 1678-7250.

Nogueira, T. S. *et. al.* (2022). Analysis of the Brazilian Artisanal Cheese Market from the Perspective of Social Networks. In: Abraham, A., Gandhi, N., Hanne, T., Hong, TP., Nogueira Rios, T., Ding, W. (eds) Intelligent Systems Design and Applications. ISDA 2021. Lecture Notes in Networks and Systems, vol 418. Springer, Cham.

Van Rossum G. *et al*. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Honnibal, M. *et al*. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Bird S. *et al*. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

Ruggiero, M. (2022). Tendências do consumo de lácteos em 2022 – como estamos até o momento? Scanntech. Fórum Milkpoint.