Mineração de dados em rede social para avaliação de tendências de consumo do queijo artesanal no Brasil

Thallys da Silva Nogueira[1], Kennya Beatriz Siqueira[2], Priscila Vanessa Zabala Capriles Goliatt[1]

[1]Programa de Pós-Graduação em Modelagem Computacional – Universidade Federal de Juiz de Fora (UFJF) – 36036-900 – Juiz de Fora – MG – Brazil

[2]Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) - Gado de Leite – 36038-330 – Juiz de Fora – MG – Brazil

{thallysnogueira,capriles}@ice.ufjf.br, kennya.siqueira@embrapa.br

Abstract. *The present work refers to the results obtained by the development of a computational system called "Consumer Observatory". Based on the concepts of business intelligence, it is capable of collecting, storing, processing and extracting information from data of the Twitter social network. The objective is to evaluate the consumption trends by identifying the characteristics and consumption habits of artisanal cheese in Brazil. Using artificial intelligence techniques, data mining and natural language processing is possible to identify information about the scenario of consumption trends for artisanal cheeses in Brazil, unknown until now.*

Resumo. *O presente trabalho refere-se aos resultados obtidos pelo desenvolvimento de um sistema computacional denominado "Observatório do Consumidor". Baseando-se nos conceitos de business intelligence é capaz de coletar, armazenar, processar e extrair informações em dados da rede social Twitter. O objetivo é avaliar tendências de consumo por meio da identificação de características e hábitos de consumo do queijo artesanal no Brasil. Fazendo-se uso de técnicas de inteligência artificial, mineração de dados e processamento de linguagem natural, é possível identificar informações sobre o cenário das tendências de consumo dos queijos artesanais no Brasil, até então desconhecidas.*

1. Introdução

Desde sempre, entender bem o mercado em que se atua, assim como o consumidor, é uma tarefa muito importante para o desenvolvimento de estratégias em ambiente empresarial. O *Business Intelligence* (BI) tem sido bastante falado nos dias atuais e refere-se a um conjunto de conceitos e metodologias que visam auxiliar e tornar mais eficiente o processo de tomada de decisão de uma empresa, transformando dados históricos e correntes em informação, e informação em decisões mais assertivas [Turban *et al.* 2010]. Com o crescente uso das redes sociais, empresas estão cada vez mais atentas ao que é exposto nestes ambientes virtuais. Segundo Wasserman e Faust (1994), as redes sociais são definidas como um conjunto composto de dois elementos que são: (i) os atores (pessoas, instituições ou grupos) e (ii) as suas conexões (interações ou laços sociais).

Com tantas funcionalidades disponíveis, as redes sociais conectam pessoas de todo o mundo e o número de pessoas que possui um cadastro nessas redes tem aumentado ano após ano [DataReportal 2021a]. Entre abril e junho de 2019, a quantidade de pessoas que possuía cadastro em alguma rede social chegou a 3,5 bilhões de pessoas. Em janeiro de 2020, atingiu-se um total de 3,8 bilhões de usuários e, em janeiro de 2021, chegou-se à marca de 4,2 bilhões de usuários [DataReportal 2021a]. Estes números podem não representar de fato contas individuais de usuários, entretanto, sugerem que mais da metade da população atual do planeta utiliza alguma rede social [Worldometers 2021].

No Brasil, a realidade é semelhante. Em janeiro de 2021, o País possuía uma população de 213,3 milhões de pessoas [DataReportal 2021b]. Cerca de 160 milhões de internautas ficavam em média 10h 08min conectados na internet através de qualquer dispositivo. Neste intervalo de tempo, gastavam, em média, 3h 42min com redes sociais, 4h 02min em frente à TV (TV aberta, *streaming* e *on demand*), 1h 52min ouvindo músicas por serviços de *streaming* e 1h 17min usando consoles de videogame. Tem-se, então, que cerca de 150 milhões de brasileiros possuem algum cadastro em rede social, totalizando 70,3% da população [DataReportal 2021a].

Com tantas redes sociais disponíveis, um grande volume de dados é gerado e, por esse motivo, o termo *Big Data* está muito presente nestes ambientes, fazendo referência ao grande volume, variedade (podendo estes dados serem mensagens de texto, comentários, vídeos, áudios, fotos) e velocidade com que são gerados. Analisar dados provenientes destes ambientes demanda novas maneiras de processar os dados e gerar informação útil, melhorando a percepção e a tomada de decisão das empresas [Gartner 2021]. Áreas como mineração de dados, inteligência artificial e ciência de dados, têm ganhado destaque por abordarem técnicas e metodologias capazes de identificar tais informações neste vasto universo de dados.

Neste âmbito, alternativas para entender o perfil do consumidor e tendências de consumo têm sido estudadas como forma de suprir as deficiências da pesquisa de mercado tradicional. Para Barabba e Zaltaman (1991), tal pesquisa é definida como o "*processo de ouvir a voz do mercado e transmitir à administração informações a respeito*", tendo como principal objetivo direcionar a tomada de decisão. Essa pesquisa de mercado tradicional é feita há décadas pelas empresas e por esse motivo já é validada e consegue atingir as expectativas das empresas. Porém, a pesquisa de mercado

tradicional pode ser ineficiente por se apresentar, em alguns casos, como questionários que podem ser extensos e com longo tempo de aplicação, o que leva a não adesão de participantes; pela necessidade de profissionais qualificados para a aplicação dos testes; pela limitação geográfica em pesquisas de extensão (inter-)continental; e principalmente pela dificuldade em se obter uma amostra bem representativa e distribuída [Veríssimo *et al*. 2018].

Pensando nisso, a Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA Gado de Leite, inovou em parceria com a 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), desenvolvendo o projeto "Observatório do Consumidor", uma plataforma que busca desenvolver alternativas às pesquisas de mercado tradicionais. Baseando-se na coleta e análise de dados de redes sociais e fazendo uso de técnicas de inteligência artificial, mineração de dados, *web*-semântica e processamento de linguagem natural, o sistema identifica o perfil dos consumidores, assim como suas tendências de consumo. A ideia é aplicar a transformação digital nas empresas, usando o grande volume de dados e informações disponíveis nas redes sociais para entender o mercado consumidor brasileiro, de modo a fornecer informações dinâmicas e em tempo real sobre as mudanças de comportamento e tendências de consumo no País.

O primeiro caso de estudo da plataforma do Observatório do Consumidor 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]. Nesses trabalhos, foi empregada a rede social Twitter [Twitter 2021], que é uma das redes sociais mais usadas no mundo com cerca de 328 milhões de contas ativas [Mohamed 2018]. Essa rede oferece aos usuários um espaço para conversação, publicação de conteúdos chamados de *tweets*, fotografias e vídeos. No contexto do Brasil, de acordo com DataReportal (2021b), cerca de 51,6% dos brasileiros entre 16 e 64 anos que usam a internet possuem uma conta no Twitter, tornando esta a sexta rede mais usada no Brasil, ficando atrás do Youtube, Facebook, Whatsapp, Instagram e FB Messenger.

Com intuito de mostrar o potencial que o Observatório do Consumidor tem para a extração de informações presentes em dados da rede social do Twitter, neste trabalho, avaliou-se em particular o queijo artesanal, um produto com forte apelo cultural no Brasil e que é responsável pela renda de muitas famílias em todo o País. Do lado do consumo, os queijos artesanais também se destacam por estarem alinhados com as principais tendências de consumo de alimentos atuais: regionalismo, volta às origens, produto natural, entre outras. Outro ponto interessante é que os queijos artesanais são muito versáteis dentro da cultura brasileira, podendo ser consumidos em diversas ocasiões, além de atender a diversos públicos, desde consumidores locais de baixa renda até famosos *chefs* de comida *gourmet*. No entanto, pouco se sabe sobre os hábitos e preferências de consumo de queijos artesanais no Brasil. Há uma carência de informações e estatísticas oficiais sobre esse mercado, que podem ser abordadas por meio do uso da *Business Intelligence,* facilitando a gestão de empresas de todos os portes desse setor.

2. Material e Métodos

A plataforma do Observatório do Consumidor possui três etapas principais: (i) coleta e armazenamento dos dados, (ii) processamento dos dados e (iii) pós-processamento dos dados e extração de informação. Inicialmente, foi necessário desenvolver uma lista contendo ao todo 179 palavras-chave (que neste caso, foram os nomes dos queijos artesanais de interesse), que foi construída com a supervisão e orientação de especialistas da Embrapa Gado de Leite e representantes da Empresa de Assistência Técnica e Extensão Territorial (EMATER).

Com a lista de palavras-chave validada, iniciou-se a etapa de coleta dos dados na rede social submetendo cada palavra-chave como parâmetro do algoritmo de coleta, armazenando cada *tweet* no banco de dados relacional desenvolvido. Com os dados devidamente armazenados, iniciou-se a etapa de processamento que padroniza os dados coletados removendo caracteres indesejados com o intuito de tornar análises futuras mais eficientes. Após esta etapa, com todos os *tweets* devidamente padronizados, iniciou-se a etapa de pós-processamento dos dados e extração de informação que realiza a análise de sentimentos dos *tweets* coletados, utilizando um aprendizado de máquina supervisionado para a classificação da polaridade dos sentimentos dos *tweets* em negativos, neutros e positivos e identificação de padrões relacionados à características do queijo artesanal e seus hábitos de consumo.

O modelo de análise de sentimentos foi desenvolvido na linguagem de programação Python [Van Rossum e Drake 2009] com a biblioteca *scikit-learn* [Pedregosa *et al.* 2011]. Para realização da tradução da linguagem humana para a linguagem de máquina, foi utilizada a abordagem *bag of words*[1] com o esquema TF-IDF [Sammut e Webb 2011], que é dada pela Frequência dos Termos - Inverso da Frequência do Documento[2]. Neste trabalho, foi proposto um classificador utilizando o *Voting Classifier* [Pedregosa *et al.* 2011] com voto majoritário a fim de combinar diversos classificadores (Regressão Logística, Naive Bayes, OneVSRest e OneVSOne) em um único classificador visando melhorar o desempenho e a precisão na classificação da polaridade dos sentimentos de cada *tweet.*

Para avaliação e verificação do desempenho do modelo, foi utilizada a técnica de validação cruzada utilizando o *K-fold* que consta em dividir o conjunto de dados de treinamento em K partes, sendo uma parte para treino e outra para teste seguindo um comportamento incremental. Utilizando a função *StratifiedKFold*, dividiu-se os dados em 10-*folds* usando o parâmetro *shuffle = true* a fim de garantir que os grupos fossem embaralhados mantendo-se sempre um percentual equivalente de dados de cada classe em cada uma das partições (tanto na de treinamento como na de teste). Para validação, cinco métricas foram implementadas que são a R^2, MSE (*Mean Squared Error*), RMSE (*Root Mean Squared Error*), MAE (*Mean Absolute Error*) e a matriz de confusão.

Para a construção do *dataset* de treinamento, foi utilizado cerca de 10,16% dos 82.959 *tweets* coletados com intuito de criar uma base própria e exclusiva do queijo artesanal com dados do próprio Twitter dado que não foi encontrado nenhum *dataset* específico na literatura para tal finalidade. Para a pré-classificação do conjunto de

[1] Vocabulário contendo todas as palavras contidas nos *tweets* de forma única.

[2] Técnica de determinação do peso que cada palavra possui em um documento.

treinamento, utilizou-se o *software* GOTIT [AI 2021], que usa redes neurais e análise semântica para a extração de informação em textos, fazendo uso de processamento de linguagem natural para a determinação dos sentimentos e polaridade presentes em uma frase. Ao todo, o *dataset* construído possui 8.432 *tweets* com 33,7% de classificações negativas, 13,5% para a classe neutra e 52,8% para a classe positiva.

Para analisar quais foram as características e hábitos de consumo mais mencionados nos *tweets* construiu-se mais duas listas de palavras-chave para a realização da mineração desses "padrões" nos dados. Com grande número de publicações coletadas, analisar de forma individual cada uma das postagens torna-se uma tarefa impraticável e por este motivo justifica-se o uso de técnicas de mineração de dados. Uma das listas de palavras-chave criadas têm a função de identificar as principais características do queijo artesanal (cor, modo de produção, sabor, textura dentre outros) e a outra, os principais hábitos de consumo (acompanhamentos, receitas culinárias, bebidas). Com o intuito de se obter informações positivas sobre o consumo dos queijos artesanais, foram selecionados todos os *tweets* classificados como positivos para esta análise de conteúdo.

3. Análise e Discussão dos Resultados

Ao todo foram coletados 82.959 *tweets* entre 30 de abril de 2020 e 18 de fevereiro de 2021, totalizando quarenta e duas semanas. Das 179 palavras-chave desenvolvidas para a coleta de dados, apenas 80 retornaram *tweets* ao longo das coletas realizadas, o que pode ser explicado pela existência de grande variedade de queijos artesanais que, muitas vezes, não são muito conhecidos pela população brasileira que utiliza o Twitter. Dentre as categorias de queijos analisados as 5 mais mencionadas pelos usuários foram **Coalho** (43.339 *tweets*), **Outros** (19.835 *tweets*), **Artesanal Paulista** (7.484 *tweets*), **Minas Artesanal** (5.687 *tweets*) e **Maturado** (2.925 *tweets*).

Após realizar o treinamento do modelo de análise de sentimentos, a Tabela 1 mostra os resultados obtidos por cada classificador.

Tabela 1 - Resultados dos classificadores. Em negrito está destacado o melhor resultado. Fonte: autores.

Modelo	Acurácia Média (%)	Intervalo de Confiança (%)
Regressão Logística	71,12	[68,22 - 74,02]
Naive Bayes	65,78	[62,49 - 68,08]
OneVSRest	70,94	[68,19 - 73,70]
OneVSOne	70,69	[68,16 - 73,23]
Voting Classifier	**71,20**	**[68,58 - 73,83]**

É possível verificar que após 10 rodadas realizadas pelo *k-fold*, o modelo proposto utilizando o *ensemble Voting Classifier* foi o que obteve o melhor resultado de acurácia média com 71,20% e um intervalo de confiança de [68,58 - 73,83]% sendo este modelo o utilizado para a realização da classificação da polaridade dos sentimentos dos *tweets* na ferramenta. Este resultado, em um primeiro momento, é bem expressivo pelo fato da base de treinamento do modelo ainda estar em desenvolvimento, possuir um

número reduzido de dados, além de possuir um desbalanceamento nas classes. Mesmo com tais limitadores, estes resultados estão bem próximos aos da literatura, como pode ser visto em Ankit (2018) em que o modelo de *ensemble* construído para classificações de sentimentos de *tweets* obteve uma acurácia média de 75,6%, sendo o tamanho da base de treinamento e o balanceamento de classes um dos principais fatores por este resultado.

Ao avaliar as métricas R^2, MSE, RMSE e MAE para o modelo *Ensemble Voting Classifier* que obteve o melhor resultado de acurácia média, foi possível observar que de um modo geral, tais métricas também obtiveram resultados bastante expressivos mesmo possuindo uma base de treinamento desbalanceada. Os resultados obtidos podem ser visualizados na Tabela 2.

Tabela 2 - Valores das métricas de validação onde R^2 é o coeficiente de determinação, MSE é o erro médio quadrático, o RMSE é a raiz do erro médio quadrático e o MAE é o erro absoluto mediano. Fonte: autores.

Métrica	Valores
R^2	0.87
MSE	0.16
RMSE	0.33
MAE	0.09

Com o modelo validado, a Figura 1 exibe o quantitativo total de classificações negativas, neutras e positivas encontradas pelo modelo.

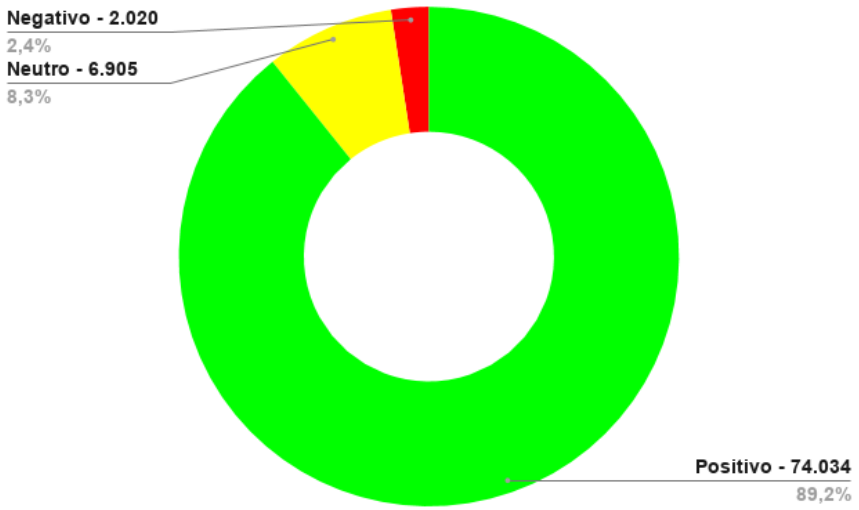

Figura 1 – Análise geral da polaridade dos sentimentos dos *tweets* sobre queijos artesanais ao longo das semanas de estudo (30/04/2020 a 18/02/2021). Fonte: autores.

De um modo geral, o sentimento positivo foi o que mais se destacou nos *tweets*, com 89,2% do total de classificações. Na maioria das vezes, a menção aos queijos artesanais nas publicações ocorreu pelo fato deste produto estar presente em diversos pratos da culinária brasileira, que são mencionados como sendo bastante saborosos.

Os *tweets* que não apresentaram adjetivos que os qualificassem tanto positivamente quanto negativamente foram classificados como neutros, e representaram cerca de 8,3%. Um exemplo sobre esta classificação seria a de um usuário somente afirmar que comeu um determinado queijo, não mencionando se gostou ou não do mesmo.

As classificações negativas apareceram em menor número representando um total de 2,4% do total. As classificações negativas no contexto dos queijos artesanais, ocorreram muito relacionadas a pessoas intolerantes à lactose. Também foi possível notar nas classificações negativas, afirmações sobre os incômodos proporcionados pelo mofo presente nos queijos e odor característico de alguns tipos de queijo.

Como mencionado anteriormente, para analisar as características e hábitos de consumo, selecionou-se todos os *tweets* com classificações positivas totalizando em 74.034 *tweets.* Ao realizar o processamento em cada *tweet*, foram identificadas cerca de 103.484 menções das características e hábitos de consumo analisadas neste estudo. Das 122 características do queijo artesanal, 98 foram encontradas e dos 104 hábitos de consumo, 103 foram retornados ao fim do processamento. Para fazer uma análise mais assertiva das tendências ao longo das semanas foi necessário analisar de forma separada cada uma das categorias devido à diferença no número de menções de atributos entre uma categoria e outra, o que dificultava muito as análises dos resultados.

Em resumo, das características do queijo artesanal, os atributos com maior número de citação foram: (i) produção **artesanal** e **tradicional**, (ii) a composição com **cebola** (sendo que esta pode não estar associada à composição do queijo) e **orégano** (presente em "dadinhos" de queijo coalho com orégano) e (iii) a categoria maturado com o termo **curado** se referindo ao processo de maturação dos queijos.

Já para os hábitos de consumo os de maior destaque foram: (i) acompanhamentos com **pão**, **carne** e **bolo**, (ii) receitas de **pão** (de/com) **queijo**, **tapioca**, ***pizza*** e **carne** (de) **sol** e (iii) bebidas, acompanhando **café** e **leite**. Com base nisso, é possível concluir que o consumo dos queijos artesanais está associado a diferentes refeições, podendo estar presente no café da manhã visto que **pão**, **café** e **leite** apareceram como acompanhamentos muito citados. Podem estar associados também a lanches, almoço e jantar com menções aos atributos **arroz**, **carne sol**, ***pizza*** (sendo um produto muito consumido no Brasil identificando a presença dos queijos artesanais como um dos ingredientes) e **tapioca** podendo esta ser consumida em qualquer refeição, inclusive como sobremesa. Isso indica que o consumo do queijo artesanal no Brasil é bastante diversificado e tem sido consumido amplamente nas mais diversas formas, refeições e acompanhamentos.

Devido a questões como características e hábitos de consumo regionais, definir uma tendência de consumo unificada para os queijos artesanais do Brasil mostrou-se muito complexo. A fim de facilitar essa identificação de tendências regionais, o Observatório do Consumidor atualmente conta com um sistema que fornece por meio de visualização gráfica a extração de informações de forma rápida e intuitiva, tornando assim a tomada de decisão local mais efetiva.

4. Considerações Finais e Indicações Futuras

Usando os "queijos artesanais" como estudo de caso, foi possível concluir que através da implementação da ferramenta do "Observatório do Consumidor" diversas

análises podem ser realizadas possibilitando identificar informações estratégicas e importantes sobre o cenário das tendências de consumo no Brasil. Através da combinação de técnicas de *Business Intelligence* e Inteligência Computacional em um só lugar, esta solução inova a maneira de se fazer pesquisas de mercado com análises das opiniões expressas pelos usuários/consumidores nas redes sociais, sem a influência e a necessidade de aplicação de questionários extensos, rompendo barreiras geográficas com maior diversidade e alcance de pessoas, mostrando quais são as principais características e hábitos/tendências de consumo. Como trabalhos futuros, pretende-se ampliar o universo de produtos e redes sociais a serem analisados, em que espera-se englobar outras categorias de derivados lácteos para a coleta, processamento e análise de dados.

References

AI, GOTIT. GOTIT - Sentiment. (2021), https://gotit.ai/en-us/Home/Sentiment.

Ankit Saleena, Nabizath. (2018). An ensemble classification system for twitter sentiment analysis. Procedia Computer Science, v. 132, p. 937–946. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science, https://www.sciencedirect.com/science/article/pii/S187705091830841X.

Barabba, T. and Zaltaman, P. (1991). "Hearing the voice of the market". Harvard Business School Press.

DataReportal. (2021). "Digital 2021 Global Digital Overview", https://datareportal.com/reports/digital-2021-global-digital-overview.

DataReportal. (2021). "Digital 2021 Brazil", https://datareportal.com/reports/digital-2021-brazil.

Gartner. (2021). "Definition of Big Data", https://www.gartner.com/en/information-technology/glossary/big-data.

Mohamed, Sinkadar. (2018). "100 Social Media Statistics You Must Know [2018] + Infographic", https://statusbrew.com/insights/social-media-statistics-2018.

Pedregosa *et al.*, (2011). Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830.

Sammut C. and Webb G.I. (2011). TF–IDF. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_832

Siqueira, Kennya B. and Nogueira, Thallys S. and Campos, Emerson W. and Soares, Nedson D. and Moraes, Emerson A. P. and Villela, Regina M. M. B. and David, José Maria N. and Goliatt, Priscila V. Z. C. (2020). "Análise exploratória da imagem dos lácteos em tempos de coronavírus. Indústria de Laticínios", n. 143, p. 64–66, ISSN 1678-7250.

Siqueira, Kennya B. and Nogueira, Thallys S. and Campos, Emerson W. and Soares, Nedson D. and Moraes, Emerson A. P. and Villela, Regina M. M. B. and David, José Maria N. and Goliatt, Priscila V. Z. C. (2020). "O impacto da pandemia no consumo de lácteos no Brasil". Indústria de Laticínios, n. 147, p. 36–38, ISSN 1678-7250.

Turban, E. and Sharda, R. and Delen, D. (2010) “Decision Support and Business Intelligence Systems”. p.720

Twitter, Inc. (2021). “Twitter”, https://twitter.com.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Veríssimo, B and Lepre, L. and Tincani, D. (2018). “Diferenças entre pesquisa de marketing e pesquisa de neuromarketing”.

Wasserman, S. and Faust, k. (1994). “Social network analysis: methods and applications”. Cambridge University Press.

Worldometers. (2021). “World Population Clock: 7,8 Billion People”, https://www.worldometers.info.