

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

Thallys da Silva Nogueira

Mineração de Dados em Rede Social para Avaliação de Tendências de  
Consumo do Queijo Artesanal no Brasil

Juiz de Fora

2021

Thallys da Silva Nogueira

Mineração de Dados em Rede Social para Avaliação de Tendências de  
Consumo do Queijo Artesanal no Brasil

Dissertação de Mestrado 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.

Orientadora: Profa. Dra. Priscila Vanessa Zabala Capriles Goliatt

Juiz de Fora

2021

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

da Silva Nogueira, Thallys.

Mineração de Dados em Rede Social para Avaliação de Tendências de Consumo do Queijo Artesanal no Brasil / Thallys da Silva Nogueira. – 2021.  
112 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt

Dissertação de Mestrado – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas / Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional , 2021.

1. Mineração de dados. 2. Aprendizado de Máquina. 3. Análise de Sentimentos. I. Capriles, Priscila, orient. II. Título.

Thallys da Silva Nogueira

Mineração de Dados em Rede Social para Avaliação de Tendências de  
Consumo do Queijo Artesanal no Brasil

Dissertação de Mestrado 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.

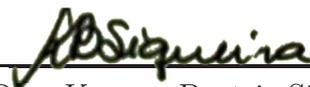
Aprovada em: 18 de Março de 2021

BANCA EXAMINADORA



---

Profa. Dra. Priscila Vanessa Zabala Capriles Goliatt  
Orientadora  
Universidade Federal de Juiz de Fora



---

Dra. Kenna Beatriz Siqueira  
Empresa Brasileira de Pesquisa Agropecuária  
(EMBRAPA) - Gado de Leite



---

Profa. Dra. Luciana Conceição Dias Campos  
Universidade Federal de Juiz de Fora

Dedico este trabalho a Deus, aos meus pais, aos meus irmãos, à minha namorada, aos meus avós em especial à Raulina Maria de Jesus que lá do alto sei que acompanha todas as minhas batalhas e conquistas. Amo todos vocês!

## AGRADECIMENTOS

A Deus por me conceder forças em toda a caminhada pois não foi fácil, houve momentos que parecia impossível chegar até aqui. Aos meus pais Carlos Roberto Nogueira (Beto) e Waldisleia Aparecida da Silva (Kiki), meus irmãos Thays da Silva Nogueira e Thomás da Silva Nogueira pelo companheirismo, minha namorada Luisa Silva Ribeiro pela paciência e apoio durante toda a minha grande jornada. Todas as palavras de incentivo, encorajamento, força nos momentos mais difíceis e broncas que me deram quando eu precisava, me moldaram e foram responsáveis por eu ser o que sou hoje, sem sombra de dúvidas foram essenciais para a finalização de mais esta etapa. Agradeço também a todos os meus familiares, que de alguma forma sempre estiveram presentes e me apoiaram em cada uma das fases de minha vida.

A todos os colegas e amigos que fiz durante os cursos de Ciências Exatas, Ciência da Computação, Engenharia Computacional, da Code Empresa Júnior de Computação, do Programa de Pós-Graduação em Modelagem Computacional e de todos os demais cursos da UFJF. Agradeço também os colegas e amigos que fiz na Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA Gado de Leite. À todos, meu sincero obrigado por todos os auxílios prestados, pelas longas conversas, histórias, risadas, reuniões e momentos que ficarão eternizados em minha memória.

À Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA Gado de Leite de Juiz de Fora, um agradecimento em especial por ter aberto as portas em minha jornada, onde tive a oportunidade de participar do Programa de Residência Zootécnica Digital sendo esta a minha primeira grande oportunidade, sendo estagiário e integrante da equipe do “Observatório do Consumidor: o caso dos queijos artesanais no Brasil” juntamente dos meus colegas estagiários Emerson W. Campos e Nedson D. Soares, aos pesquisadores da EMBRAPA Fábio Homero Diniz, Maria de Fátima Ávila Pires, Nívea Maria Vicentini, à pesquisadora, supervisora e uma das idealizadoras do projeto na EMBRAPA Kenna Beatriz Siqueira pela orientação, confiança e incentivo ao longo de todo o projeto. Aos professores Emerson A. P. Moraes, José Maria N. David, Regina M.M.B Villela pelas dicas e orientações aos meus colegas estagiários e a mim. Um agradecimento especial à minha orientadora Priscila Vanessa Zabala Capriles Goliatt pela orientação, amizade, incentivo e principalmente, pela paciência e pela confiança em mim depositada, sem a qual este trabalho não se realizaria.

Agradeço também à Prefeitura de Ewbank da Câmara - MG de minha cidade natal por todo apoio e incentivo, pelo longo período que forneceram auxílios para transporte que possibilitaram à nós estudantes, o deslocamento até à universidade, isso ajudou e muito nesta minha conquista.

Por fim, agradeço à Universidade Federal de Juiz de Fora e a Empresa Brasileira

de Pesquisa Agropecuária - EMBRAPA Gado de Leite por me fornecer a estrutura necessária para o meu crescimento pessoal, formação acadêmica e profissional proporcionando momentos inesquecíveis.

“Tente uma, duas, três vezes e se possível tente a quarta, a quinta e quantas vezes for necessário. Só não desista nas primeiras tentativas, a persistência é amiga da conquista. Se você quiser chegar aonde a maioria não chegou, faça aquilo que a maioria não faz.”

Bill Gates.

## RESUMO

O desenvolvimento de alternativas à pesquisa de mercado tradicional tem sido um grande desafio em ambientes corporativos. Tais pesquisas, têm embasamento científico, são bem validadas e trazem bons resultados, porém, fatores como limitações geográficas, longo tempo de aplicação e análise, alto custo, além de possuírem formulários que podem ser extensos e muitas vezes com baixa adesão podem inviabilizar seu uso. Com tantas pessoas utilizando as mais diversas redes sociais, as empresas estão cada vez mais atentas ao que se comenta nesses ambientes virtuais. Com tantas mensagens, *posts*, comentários e áudios, um grande volume de dados é gerado e, por isso, o termo *Big Data* neste ambiente está muito presente se referindo ao grande volume, variedade e velocidade com que os dados são gerados. Conhecer o perfil do consumidor e as tendências de consumo pode ajudar muito na tomada de decisões estratégicas nos mais diversos setores e segmentos comerciais. Dito isso, o presente trabalho refere-se ao desenvolvimento de um sistema computacional capaz de coletar, armazenar, processar e extrair informações em dados da rede social Twitter, com o objetivo de avaliar tendências de consumo por meio da identificação de características e hábitos de consumo do queijo artesanal no Brasil. Para isso, inicialmente foi desenvolvido um conjunto de palavras-chave referentes aos nomes dos queijos artesanais de interesse e um algoritmo capaz de coletar esses dados no Twitter. Em seguida, foi necessário construir um modelo de banco de dados capaz de armazenar com eficiência os dados recém-coletados e os resultados de outras análises. Para identificar a polaridade dos sentimentos presentes nos *tweets* desenvolveu-se um modelo utilizando o **Ensemble Voting Classifier** para a realização da classificação das publicações nas classes negativo, neutro e positivo. Foi necessária a criação de um conjunto de dados de treinamento específico para queijos artesanais, uma vez que não foi encontrado na literatura nenhum que atendesse a essas especificações. Utilizou-se técnicas de Processamento de Linguagem Natural para tratamento dos dados textuais e extração de informação dos mesmos. Foi desenvolvido um conjunto de palavras-chave baseada na opinião de pesquisadores da EMBRAPA Gado de Leite, representantes da Empresa de Assistência Técnica e Extensão Territorial (EMATER) e análise das palavras mais frequentes nos *tweets* sobre queijo artesanal, capazes de identificar as características e os hábitos de consumo do queijo artesanal. Com a análise da opinião expressa pelo consumidor na rede social do Twitter sobre os queijos artesanais no Brasil, verificou-se que definir uma tendência geral para todos os queijos artesanais foi uma tarefa impraticável devido a questões como o regionalismo, características e hábitos de consumo que variam para cada tipo de queijo, fato este que pôde ser observado pela análise do perfil de consumo do queijo Coalho. Entretanto, levando em consideração os resultados obtidos, o sistema computacional desenvolvido mostrou-se ser capaz de combinar várias técnicas e metodologias em um só lugar, tornando a tomada de decisão mais eficiente pautada pela análise de dados reais, inovando assim a forma de

se fazer pesquisas de mercado tradicionais.

Palavras-chave: Mineração de dados. Aprendizagem de Máquina. Análise de Sentimentos.  
Redes Sociais. Queijo Artesanal.

## ABSTRACT

The development of alternatives to traditional market research has been a major challenge in corporate environments. Such research has a scientific basis, is well validated and brings good results, however, factors such as geographic limitations, long time of application and analysis, high cost, in addition to having forms that can be extensive and often with low adherence can make their use unfeasible. With so many people using the most diverse social networks, companies are increasingly attentive to what is commented on in these virtual environments. With so many messages, posts, comments and audios, a large volume of data is generated and, therefore, the term Big Data in this environment is very present referring to the large volume, variety and speed with which the data are generated. Knowing the consumer profile and consumption trends can help a lot in making strategic decisions in the most diverse sectors and commercial segments. That said, the present work refers to the development of a computational system capable of collecting, storing, processing and extracting information in data from the social network Twitter, with the objective of evaluating consumption trends through the identification of consumption characteristics and habits of artisanal cheese in Brazil. For this, initially a set of keywords was developed referring to the names of the artisanal cheeses of interest and an algorithm capable of collecting this data on Twitter. Then, it was necessary to build a database model capable of efficiently storing the newly collected data and the results of other analyzes. To identify the polarity of the feelings present in the tweets, a model was developed using the Voting Classifier Ensemble to perform the classification of publications in the negative, neutral and positive classes. It was necessary to create a specific training data set for artisanal cheeses, since no literature that met these specifications was found in the literature. Natural Language Processing techniques were used to treat textual data and extract information from them. A set of keywords was developed based on the opinion of researchers at EMBRAPA Dairy Cattle, representatives of the Assistance Technical and Territorial Extension Company (EMATER) and the analysis of the most frequent words in tweets about artisanal cheese, capable of identifying the characteristics and consumption habits of the artisanal cheese. With the analysis of the opinion expressed by the consumer on Twitter's social network about artisanal cheeses in Brazil, it was found that defining a general trend for all artisanal cheeses was a practical task due to issues such as regionalism, characteristics and consumption habits vary for each type of cheese, a fact that could be observed by analyzing the consumption profile of Coalho cheese. However, taking into account the results obtained, the computer system developed proved to be capable of combining several techniques and methodologies in one place, making decision-making more efficient based on the analysis of real data, thus innovating the way of doing traditional market research.

Keywords: Data Mining. Machine Learning. Sentiment Analysis. Social Networks.

Artisanal Cheese.

## LISTA DE ILUSTRAÇÕES

Figura 1 – Mapa de queijos artesanais . . . . .	24
Figura 2 – Interesse pelos termos de busca <i>lanches</i> , <i>receitas</i> e <i>delivery</i> nos últimos 5 anos no Brasil . . . . .	31
Figura 3 – Interesse pelo termo de pesquisa <i>Inteligência Artificial</i> no Brasil	35
Figura 4 – Hierarquia das áreas de estudo . . . . .	36
Figura 5 – Exemplo de aprendizado supervisionado . . . . .	37
Figura 6 – Exemplo de aprendizado não-supervisionado . . . . .	38
Figura 7 – Comparação entre algoritmos de aprendizado de máquina tradicionais e de aprendizagem profunda . . . . .	39
Figura 8 – Representação esquemática do fluxo de tarefas realizado neste trabalho	46
Figura 9 – Quantidade de palavras-chave por categoria de queijo . . . . .	47
Figura 10 – Modelo do Banco de Dados Relacional da Plataforma do “Observatório do Consumidor” . . . . .	49
Figura 11 – Desbalanceamento presente nos dados da base de treinamento do modelo . . . . .	54
Figura 12 – Representação esquemática do diagrama de validação cruzada com $k = 5$ . . . . .	55
Figura 13 – Matriz de Confusão . . . . .	56
Figura 14 – Nuvem das palavras mais citadas ao longo das semanas de estudo	58
Figura 15 – Quantidade de <i>tweets</i> ao longo das semanas . . . . .	61
Figura 16 – Categorias de queijos artesanais mais citados no Brasil no período da pesquisa . . . . .	62
Figura 17 – Publicações por estados brasileiros sobre queijo artesanal ao longo das semanas de estudo . . . . .	63
Figura 18 – Resultados da Matriz de Confusão . . . . .	66
Figura 19 – Análise dos conteúdos ao longo do tempo . . . . .	68
Figura 20 – Análise de conteúdo da categoria <i>Tamanho</i> nos <i>tweets</i> sobre queijos artesanais . . . . .	70
Figura 21 – Análise de conteúdo da categoria <i>Odor</i> nos <i>tweets</i> sobre queijos artesanais . . . . .	71
Figura 22 – Análise de conteúdo da categoria <i>Fungo</i> nos <i>tweets</i> sobre queijos artesanais . . . . .	71
Figura 23 – Análise de conteúdo da categoria <i>Selo Arte</i> nos <i>tweets</i> sobre queijos artesanais . . . . .	72
Figura 24 – Análise de conteúdo da categoria <i>Preço</i> nos <i>tweets</i> sobre queijos artesanais . . . . .	73

Figura 25 – Análise de conteúdo da categoria <b>Textura</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	74
Figura 26 – Análise de conteúdo da categoria <b>Cor</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	74
Figura 27 – Análise de conteúdo da categoria <b>Embalagem</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	75
Figura 28 – Análise de conteúdo da categoria <b>Sabor</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	76
Figura 29 – Análise de conteúdo da categoria <b>Maturado</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	76
Figura 30 – Análise de conteúdo da categoria <b>Composição</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	77
Figura 31 – Análise de conteúdo da categoria <b>Produção</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	78
Figura 32 – Análise de conteúdo da categoria <b>Bebidas</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	79
Figura 33 – Análise de conteúdo da categoria <b>Receitas</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	81
Figura 34 – Análise de conteúdo da categoria <b>Acompanhamentos</b> nos <i>tweets</i> sobre queijos artesanais . . . . .	82
Figura 35 – Análise geral da polaridade dos sentimentos dos <i>tweets</i> sobre queijos artesanais . . . . .	83
Figura 36 – Análise da polaridade dos sentimentos ao longo das semanas de estudo	84
Figura 37 – Nuvem de palavras referente à semana 4 (21-05-2020 a 28-05-2020)	85
Figura 38 – Nuvem de palavras referente à semana 9 (25-06-2020 a 2-07-2020)	86
Figura 39 – Nuvem de palavras referente à semana 19 (03-09-2020 a 10-09-2020)	87
Figura 40 – Análise da quantidade de menções aos termos das categorias <b>Consumidores reais</b> e <b>Potenciais consumidores</b> referentes ao ato de consumir queijos artesanais em conjunto com seus acompanhamentos e receitas	89
Figura 41 – Regiões do Brasil que mencionaram o Queijo Coalho . . . . .	91
Figura 42 – Tela inicial da plataforma do Observatório do Consumidor . . . . .	93
Figura 43 – Tela com os resultados da análise de sentimentos e análise temporal da quantidade de <i>tweets</i> ao longo do tempo . . . . .	94
Figura 44 – Tela com os resultados das características e hábitos de consumo do queijo artesanal . . . . .	95

## LISTA DE TABELAS

Tabela 1 – Características e hábitos de consumo dos queijos artesanais analisados nos <i>tweets</i> . . . . .	59
Tabela 2 – Características do consumo analisados nos <i>tweets</i> . . . . .	60
Tabela 3 – Resultados dos classificadores. . . . .	64
Tabela 4 – Valores das métricas de validação . . . . .	65
Tabela 5 – Métricas de Avaliação para o Modelo <b>Voting Classifier</b> . . .	67
Tabela 6 – Lista de palavras-chave usadas para coleta no Twitter . . . . .	105

## LISTA DE ABREVIATURAS E SIGLAS

BI	<i>Business Intelligence</i>
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
UFJF	Universidade Federal de Juiz de Fora
EMATER	Empresa de Assistência Técnica e Extensão Territorial
COMERQUEIJO	Associação de Comerciantes de Queijos Artesanais Brasileiros
QMA	Queijo Minas Artesanal
APROALAGOA	Associação dos Produtores de Queijo Artesanal de Alagoa
ABIQ	Associação Brasileira das Indústrias de Queijo
IBGE	Instituto Brasileiro de Geografia e Estatística
IA	Inteligência Artificial
PLN	Processamento de Linguagem Natural
TF-IDF	Frequência dos Termos - Inverso da Frequência do Documento
SGBD	Sistema de Gerenciamento de Banco de Dados
$R^2$	Coefficiente de Determinação
MSE	Erro Médio Quadrático
RMSE	Raiz do Erro Médio Quadrático
MAE	Erro Absoluto Mediano
API	<i>Application Programming Interface</i>

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>18</b>
1.1	APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÃO DO PROBLEMA	18
1.2	OBJETIVOS	20
1.2.1	Objetivo Geral	20
1.2.2	Objetivos Específicos	20
1.3	ORGANIZAÇÃO DO TRABALHO	20
<b>2</b>	<b>REVISÃO DA LITERATURA</b>	<b>22</b>
2.1	QUEIJO ARTESANAL	22
2.1.1	Características dos Tipos de Queijos Artesanais	26
2.1.1.1	Minas Artesanal	26
2.1.1.2	Artesanal Paulista	26
2.1.1.3	Alagoa	27
2.1.1.4	Coalho	27
2.1.1.5	Colonial	28
2.1.1.6	Manteiga	28
2.1.1.7	Mantiqueira de Minas	28
2.1.1.8	Requeijão	28
2.1.1.9	Serrano	28
2.1.1.10	Cabacinha	29
2.1.1.11	Contestado	29
2.1.1.12	Kochkäse	29
2.1.1.13	Marajó	29
2.1.1.14	Nicola	30
2.1.1.15	Porongo	30
2.1.1.16	Queijinho	30
2.1.1.17	Maturado	30
2.1.1.18	Outros	30
2.1.2	Efeitos da Pandemia no Consumo	30
2.2	TENDÊNCIAS DE CONSUMO DE ALIMENTOS	32
2.3	BUSINESS INTELLIGENCE	33
2.4	INTELIGÊNCIA ARTIFICIAL	34
2.4.1	Aprendizado de Máquina	35
2.4.1.1	Aprendizado Supervisionado	36
2.4.1.2	Aprendizado Não-supervisionado	37
2.4.2	Aprendizagem Profunda	38
2.5	MINERAÇÃO DE DADOS	38

2.5.1	Processamento de Linguagem Natural (PLN) . . . . .	40
2.5.1.1	Tratamento dos dados . . . . .	40
2.5.1.1.1	Tokenização . . . . .	40
2.5.1.1.2	Remoção de <i>stopwords</i> . . . . .	41
2.5.1.1.3	Stemming . . . . .	42
2.5.1.1.4	Vetorização TF-IDF . . . . .	42
2.5.1.1.5	<i>N</i> -gramas . . . . .	43
2.6	ANÁLISE DE SENTIMENTOS . . . . .	44
<b>3</b>	<b>MATERIAL E MÉTODOS . . . . .</b>	<b>46</b>
3.1	FLUXO DE TAREFAS . . . . .	46
3.1.1	Etapa 1: Coleta e Armazenamento dos Dados . . . . .	47
3.1.1.1	Determinação do Conjunto de Busca . . . . .	47
3.1.1.2	Coleta de Dados do Twitter . . . . .	48
3.1.1.3	Estrutura dos dados do Twitter ( <i>tweet</i> ) . . . . .	48
3.1.1.4	Armazenamento dos Dados . . . . .	48
3.1.1.4.1	Modelagem do Banco de Dados MySQL . . . . .	49
3.1.2	Etapa 2: Processamento dos Dados . . . . .	50
3.1.2.1	Mineração de Dados Textuais . . . . .	50
3.1.3	Etapa 3: Pós-Processamento dos Dados . . . . .	50
3.1.3.1	Interpretação dos Dados . . . . .	50
3.2	MODELO DE ANÁLISE DE SENTIMENTOS . . . . .	50
3.2.1	Ensemble Voting Classifier . . . . .	51
3.2.1.1	Classificadores . . . . .	51
3.2.1.1.1	Regressão Logística . . . . .	51
3.2.1.1.2	Naïve Bayes . . . . .	51
3.2.1.1.3	OneVsRest e OneVsOne . . . . .	52
3.2.1.2	Votação por Decisão da Maioria . . . . .	53
3.2.2	Base de Treinamento do Modelo . . . . .	53
3.2.3	Validação Cruzada . . . . .	54
3.2.4	Métricas de Avaliação . . . . .	55
3.3	ANÁLISE DE CONTEÚDO . . . . .	57
3.3.1	Análise das Características e Hábitos de Consumo do Queijo Artesanal . . . . .	57
3.3.2	Análise de Consumidores Reais e Potenciais Consumidores dos Queijos Artesanais . . . . .	60
<b>4</b>	<b>RESULTADOS E DISCUSSÃO . . . . .</b>	<b>61</b>
4.1	PANORAMA GERAL . . . . .	61
4.1.1	Análise Temporal . . . . .	61
4.1.2	Categorias de Tipos de Queijos Artesanais mais Citados no Twitter . . . . .	62
4.1.3	Estados Brasileiros que mais Comentaram Sobre Queijos Artesanais . . . . .	63

4.2	MODELO DE ANÁLISE DE SENTIMENTOS . . . . .	64
4.2.1	Validação Cruzada . . . . .	64
4.2.2	Valores das Métricas de Validação . . . . .	65
4.2.3	Matriz de Confusão . . . . .	65
4.3	ANÁLISE DE CONTEÚDO . . . . .	67
4.3.0.1	Características e Hábitos de Consumo do Queijo Artesanal . . . . .	69
4.3.1	Análise da Polaridade dos Sentimentos e Consumo ao Longo das Semanas	83
4.3.2	Análise de Consumidores Reais e Potenciais Consumidores dos Queijos Artesanais . . . . .	88
4.3.2.1	Análise do Perfil de Consumo do Queijo Coalho . . . . .	90
4.4	TELAS DA PLATAFORMA DO OBSERVATÓRIO DO CONSUMIDOR	92
5	<b>CONCLUSÃO . . . . .</b>	<b>96</b>
6	<b>TRABALHOS FUTUROS . . . . .</b>	<b>98</b>
	<b>REFERÊNCIAS . . . . .</b>	<b>99</b>
	<b>APÊNDICE A – LISTA DE PALAVRAS-CHAVE . . . . .</b>	<b>105</b>
	<b>APÊNDICE B – DEFINIÇÕES DOS ATRIBUTOS . . . . .</b>	<b>111</b>

# 1 INTRODUÇÃO

## 1.1 APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÃO DO PROBLEMA

Desde sempre, a comunicação possui um papel muito importante na forma como os seres humanos se relacionam. Ao longo da evolução humana, diversas maneiras de se comunicar foram desenvolvidas até os dias atuais. As redes sociais virtuais são um bom exemplo dessa evolução. Segundo Wasserman e Faust (1994), uma rede social é definida 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 atualmente abrangem uma enorme quantidade de pessoas conectadas que, segundo relatório publicado por *Global Digital Statshot* entre abril e junho de 2019 (KEMP, 2019), o número de pessoas que possuíam um cadastro em alguma rede social chegou a 3,5 bilhões de pessoas, sendo quase a metade da população atual do planeta com cerca de 7,6 bilhões de pessoas (WORLDMETERS, 2017).

No Brasil não é diferente, segundo relatório do site *DataReportal* (KEMP, 2020), em janeiro de 2020, cerca de 150,4 milhões de internautas ficavam em média cerca de 9h 17min conectados na internet através de qualquer dispositivo. Neste intervalo de tempo, gastavam cerca de 3h 31min com redes sociais, 3h 51min em frente à TV (TV aberta, streaming e *on demand*), 1h 41min escutando músicas por serviços de *streaming* e 1h 14min usando consoles de videogame. Ainda com base neste relatório, cerca de 140 milhões de brasileiros possuem algum cadastro em rede social, totalizando 66% da população.

Com tantas pessoas usando as mais diversas redes sociais, empresas estão cada vez mais atentas ao que é exposto nestes ambientes virtuais. Com tantas mensagens, postagens, comentários e áudios, um grande volume de dados é gerado e, por esse motivo, o termo *Big Data* neste ambiente está muito presente fazendo referência ao grande volume, variedade e velocidade com que os dados são gerados. Isso demanda novas maneiras de processar os dados e gerar informação útil, melhorando a percepção e a tomada de decisão (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. Assim, atualmente os dados são peças chave para a geração de informação e de acordo com o pesquisador David Buckingham da Loughborough University, "*Dado é o novo petróleo! Precisamos encontrá-lo, extraí-lo, refiná-lo, distribuí-lo e monetizá-lo*".

No ambiente empresarial o *Big Data* é estratégico. Neste contexto, o *Business Intelligence* (BI), também conhecido como Inteligência de Negócios ou Inteligência Empresarial, é um conjunto de conceitos e metodologias que visam auxiliar e tornar mais eficaz a tomada de decisão em diferentes situações no ambiente corporativo, fazendo uso de dados históricos e correntes aumentando as chances de uma decisão ser tomada de forma correta,

transformando dados em informação e informação em decisões, até de fato gerar a ação (TURBAN et al., 2010). Neste âmbito, o universo de dados disponíveis é vasto e tem sido muito utilizado para entender o que os consumidores comentem sobre um determinado produto, que configura-se em uma informação valiosa e que pode mudar estratégias de ação de uma empresa.

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 apropriada informações a respeito”, tendo como principal objetivo alertar e indicar possíveis tomadas de decisão. Essa pesquisa tradicional é feita a décadas pelas empresas e por esse motivo já é validada e consegue atingir as expectativas de quem faz o uso dela. Porém, essa mesma pesquisa de mercado tradicional pode ser ineficiente por se apresentarem, em alguns casos, como questionários 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: o caso dos queijos artesanais do Brasil”, uma plataforma que busca desenvolver alternativas às pesquisas de mercado tradicionais, baseando-se na coleta e análise de dados de redes sociais fazendo-se uso de técnicas de inteligência artificial, mineração de dados, *web-semântica* e processamento de linguagem natural, para identificar o perfil dos consumidores assim como suas tendências de consumo.

O primeiro caso de estudo da plataforma do Observatório do Consumidor foi relacionada 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 Kemp (2020), cerca de 48% 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.

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. A

sua história é bastante extensa e como forma de valorizar este produto e todo o processo de fabricação manual empregado ao longo dos anos, em 2019 o Governo Brasileiro desenvolveu a Lei Nº 13.860, que caracteriza o queijo artesanal como sendo aquele confeccionado através de métodos tradicionais, vinculando a valorização territorial, regional ou cultural, seguindo um protocolo de fabricação específico para cada tipo e variedade de queijos, e empregando boas práticas agropecuárias e de fabricação (BRASIL, 2019). Outro fator que foi levado em consideração para a escolha deste produto neste trabalho se deu pela carência de maiores informações sobre seu consumo no Brasil.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo Geral

O presente trabalho tem como objetivo geral desenvolver um sistema computacional baseado nos conceitos de *Business Intelligence* capaz de coletar, armazenar, processar e extrair informações em dados da rede social do Twitter, a fim de avaliar as tendências de consumo do queijo artesanal no Brasil, identificando quais são suas características e hábitos de consumo de forma automatizada.

### 1.2.2 Objetivos Específicos

- Desenvolver um sistema de mineração de dados para a rede social Twitter;
- Desenvolver um modelo de banco de dados relacional para armazenamento dos dados coletados;
- Desenvolver análises exploratórias dos dados coletados e armazenados;
- Implementar algoritmos de processamento de linguagem natural (PLN) para tratamento e tradução da linguagem humana para a linguagem de máquina;
- Desenvolver um modelo de Análise de Sentimentos específico para a identificação da polaridade dos sentimentos presentes nos conteúdos dos *tweets* relacionados a queijos artesanais;
- Desenvolver algoritmos de mineração de dados para identificação de características e hábitos de consumo do queijo artesanal presente nos *tweets*;
- Analisar os resultados dos diferentes perfis de consumidores dentro do contexto dos queijos artesanais no Brasil.

## 1.3 ORGANIZAÇÃO DO TRABALHO

O texto deste trabalho está organizado da seguinte forma:

- Capítulo 2 - Revisão da Literatura: neste capítulo é mostrado um breve contexto histórico sobre os queijos artesanais no Brasil juntamente com suas características, uma visão geral sobre o consumo de queijos no Brasil e o impacto da pandemia da COVID-19 nesse mercado. Em seguida faz-se um breve resumo sobre as tendências de consumo de alimentos e suas principais categorias. Em seguida, é apresentado um panorama geral sobre a inteligência de negócios e como o uso de tal metodologia pode tornar a tomada de decisões em ambiente corporativo mais assertivo. Mais adiante, é introduzida a seção de inteligência artificial exibindo as principais características abordadas por ela. E para finalizar este capítulo é abordado nas duas últimas seções os temas de mineração de dados e análise de sentimentos, onde o primeiro refere-se ao processo de preparação, limpeza e extração de informação e o segundo sobre a técnica de aprendizado de máquina por abordagem supervisionada para classificação da polaridade presente em um determinado texto.
- Capítulo 3 - Material e Métodos: neste capítulo inicialmente é mostrado o fluxo de tarefas desenvolvido no projeto. A primeira etapa se refere à coleta e armazenamento dos dados determinando qual é o conjunto de busca a ser utilizado na coleta de dados da rede social do Twitter e como foi desenvolvido o modelo de banco de dados relacional para armazenamento dos mesmos. Em seguida, a segunda etapa é referente à etapa de processamento de dados, sendo esta responsável pela mineração dos dados textuais coletados a fim de identificar informações sobre os queijos artesanais nos *tweets*. A terceira e última etapa, é referente à interpretação dos resultados obtidos. Em seguida, é mostrado em detalhes as características do modelo desenvolvido de análise da polaridade dos sentimentos dos *tweets*, mostrando detalhes sobre o método de *Ensemble Voting Classifier* e quais classificadores foram utilizados para sua construção. Em seguida, são apresentadas as técnicas de validação utilizadas para o modelo de análise da polaridade de sentimentos presentes nos *tweets* e por fim os termos chave analisados nos textos para a realização da análise de conteúdo;
- Capítulo 4 - Resultados e Discussão: neste capítulo são apresentados os resultados obtidos juntamente com suas respectivas discussões;
- Capítulo 5 - Conclusão: neste capítulo é apresentado a conclusão obtida através das análises e discussões dos resultados;
- Capítulo 6 - Trabalhos Futuros: neste capítulo são apontadas perspectivas para melhoria e construção de novas análises, expandindo-as para outras redes sociais a fim de obter um panorama mais diversificado sobre tendências de consumo.

## 2 REVISÃO DA LITERATURA

### 2.1 QUEIJO ARTESANAL

O Brasil é um país rico em diversidade e carrega consigo atualmente 520 anos de muitas histórias. O Descobrimento do Brasil, é marcado com a chegada da frota portuguesa comandada por Pedro Álvares Cabral que desembarca em um local rico em fauna e flora mas sem nenhuma das mordomias encontradas e vividas em terras portuguesas. Encontram um povo nativo que vivia exclusivamente da caça, da pesca, do extrativismo e da produção de frutos, raízes e tubérculos. Para encontrar o Brasil, Pedro Álvares Cabral e sua tripulação gastaram em média cerca de 44 dias e para que sobrevivessem a este longo trajeto, o queijo era um dos alimentos fornecidos para sua tripulação (BERNARDO, 2020).

De acordo com a carta de Pero Vaz de Caminha em 1º de maio de 1500, os índios *‘não lavram, nem criam, nem há boi, nem vaca, nem cabra, nem ovelha, nem galinha, nem outra nenhuma alimária, que costumada seja ao viver dos homens’*. Dessa forma, alimentos que lá em Portugal já eram completamente difundidos a exemplo do queijo, aqui no Brasil por mais rudimentar que fosse, tal produção não era possível, pois não havia animais que tornassem possíveis essas práticas e produções (BOARI, 2017).

Com o estabelecimento da colônia brasileira, os principais objetivos da metrópole foram basicamente extrativistas e monocultores, sendo caracterizados pelos ciclos econômicos da cana-de-açúcar e mineração. À medida que as explorações foram acontecendo, iniciou-se o processo de criação de animais como cabras, ovelhas e as vacas, primeiramente como modo de subsistência até assumir um papel relevante na geração de renda. Data-se que em 1.534 as primeiras vacas foram trazidas ao Brasil por Martim Afonso de Souza e sua esposa Ana Pimentel de Souza. Inicialmente, como não havia produção de queijo no Brasil, esses animais foram utilizados como geração de força motriz nos engenhos de trapiche e para transporte nos carros de boi (BOARI, 2017).

Passado o ciclo da cana-de-açúcar, começa então o ciclo da mineração e neste momento da história do Brasil por volta do século XVII, inicia-se a produção de queijos de forma mais acentuada. Neste período, a produção de queijos ocorria nas regiões dos atuais estados de Goiás, Mato Grosso, Bahia e especialmente em Minas Gerais (BOARI, 2017). Os anos foram passando, técnicas foram sendo passadas de geração a geração e hoje o queijo no Brasil possui um enorme destaque sendo um alimento bastante consumido nas mais diferentes receitas culinárias atrelado a um forte apelo cultural sendo a principal fonte de renda nos dias atuais de muitas famílias em todo o Brasil (PIRES et al., 1994; CAVALCANTE, 2005).

Por ser um produto regional, um desejo dos produtores de queijo artesanal era o de poder comercializar esse produto em diferentes partes do estado e do País de forma

regulamentada. Com isso, o decreto assinado no dia 18 de julho de 2019 da Lei Nº 13.860, regulamenta o Selo Arte, que permite a esses produtores venderem produtos alimentícios como queijos, mel e embutidos de maneira interestadual o que aumenta muito o mercado e a renda dessas pessoas retirando esses comerciantes da clandestinidade.

A maneira artesanal de se produzir este produto se espalhou por todo o País e em cada região brasileira um tipo de queijo específico foi se destacando e ganhando nome ao longo dos anos. Dessa forma, a Figura 1 desenvolvida pela *Slow Food* em 2018 (FOOD, 2018) mostra alguns dos queijos mais conhecidos e tradicionais de cada região do Brasil.

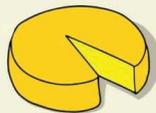
Um bioma, de acordo com o IBGE (2021) é definido como “um conjunto de vida vegetal e animal, constituído pelo agrupamento de tipos de vegetação contíguos e que podem ser identificados a nível regional, com condições de geologia e clima semelhantes e que, historicamente, sofreram os mesmos processos de formação da paisagem, resultando em uma diversidade de flora e fauna própria”. Dessa maneira no Brasil, existem 6 biomas que abrangem todas as regiões do País que são: a Amazônia, o Pantanal, a Caatinga, o Cerrado, a Mata Atlântica e o Pampas.

O bioma da Amazônia abrange os estados do Acre, Amazonas, Roraima, Rondônia, Pará, Amapá, Maranhão, Mato Grosso e Tocantins e nestes estados destacam-se 4 principais tipos de queijo que são o Queijo Manteiga, Queijo de Coalho, Queijo do Marajó e Queijo do Norte.

O bioma do Pantanal abrange os estados brasileiros do Mato Grosso e Mato Grosso do Sul, e seu restante se estende pela Bolívia e Paraguai. Neste bioma, os Queijos Nicola e Curado são os mais conhecidos.

O bioma da Caatinga abrange os estados do Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia e uma pequena área de Minas Gerais. Nesta região destacam-se 13 principais tipos de queijo, sendo os mais conhecidos os de Coalho, Queijo Creme e Queijo Manteiga. Por mais que os queijos de Coalho possuam o mesmo nome, suas características podem ser bastante distintas devido ao modo de fabricação de região para região.

O bioma do Cerrado abrange os estados de Goiás, Mato Grosso, Mato Grosso do Sul, Distrito Federal, Tocantins, Maranhão, Piauí, Bahia, Minas Gerais, São Paulo e Paraná, abrangendo ainda uma parte dos estados de Pará, Rondônia, Roraima e Amapá. Nestas regiões destacam-se o Requeijão Moreno, Queijo Cabacinha ou Flamengo, Queijo Minas Alto Paranaíba, Queijo Minas Salitre, Queijo Minas Triângulo Mineiro, Queijo Minas Araxá e o Requeijão.



# Queijos Artesanais do Brasil

regiões tradicionais

- Legenda Biomas do Brasil
- Amazônia
  - Cerrado
  - Mata Atlântica
  - Caatinga
  - Pantanal
  - Pampas

**queijo artesanal é leite cru**

Os queijos de **leite cru** possuem texturas, sabores e aromas únicos, resultado da sua rica **microbiota natural** que garantem a **particularidade** e as **qualidades benéficas** destes alimentos. Um ambiente de produção saudável, com animais vivendo e se alimentando de forma natural, influencia o equilíbrio desta microflora e interfere positivamente na **qualidade do leite**. Estes benefícios são perdidos com o processo de **pasteurização**.

O Slow Food convida a todos para agirem em defesa dos **queijos artesanais de leite cru** que representam a diversidade dos alimentos regionais, o equilíbrio com a natureza e a sabedoria das comunidades guardiãs deste patrimônio. A luta por um **alimento bom, limpo e justo** é uma luta pelo sabor, pela biodiversidade e pelo direito a uma alimentação saudável e prazerosa.

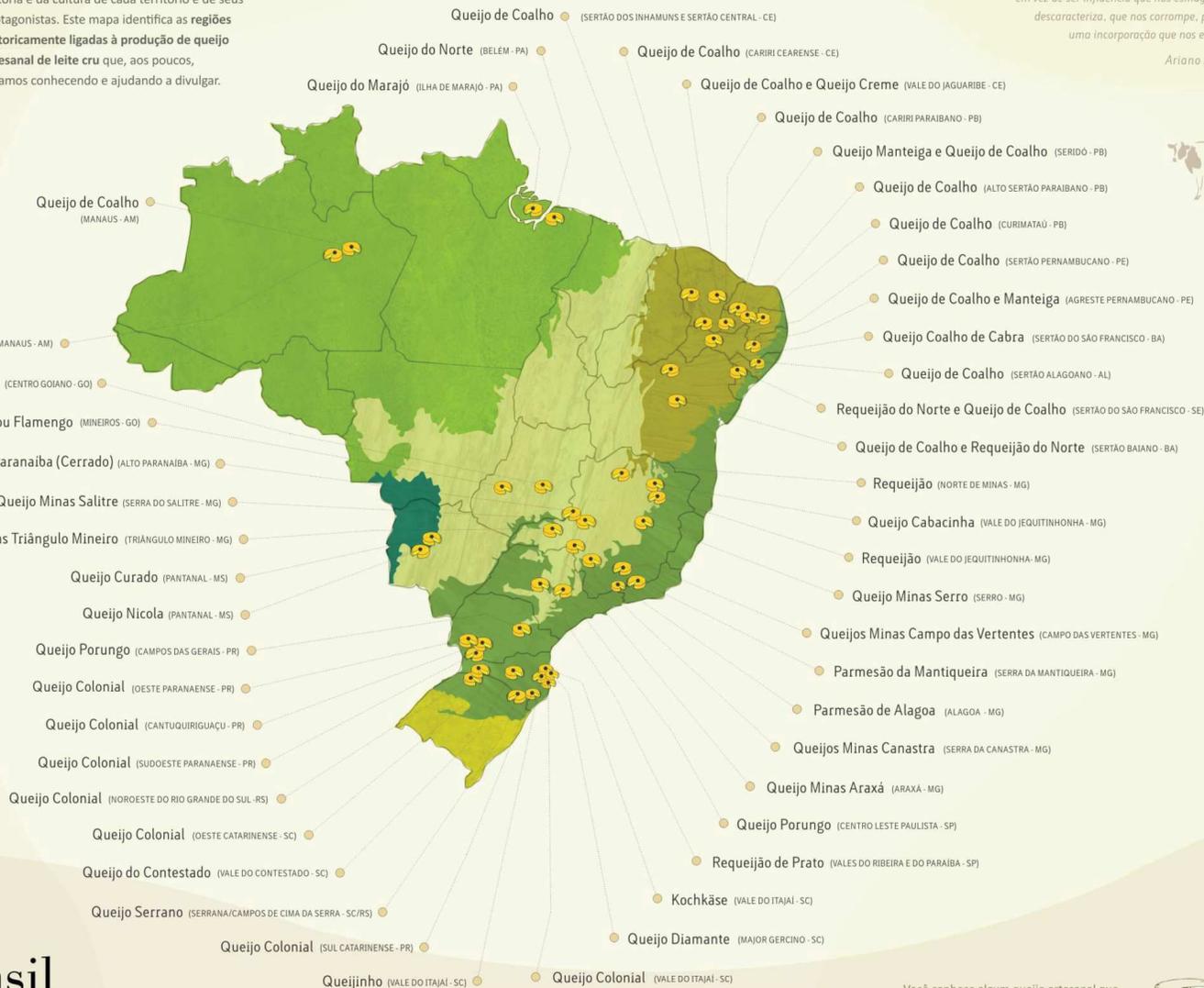
Junte-se à nós!



**Slow Food Brasil**

www.slowfoodbrasil.com

Existe uma grande variedade de **queijos artesanais** no Brasil, frutos do saber fazer tradicional, da história e da cultura de cada território e de seus protagonistas. Este mapa identifica as **regiões historicamente ligadas à produção de queijo artesanal de leite cru** que, aos poucos, estamos conhecendo e ajudando a divulgar.



«Não tenho nada contra ninguém, O que eu quero é fortalecer a nossa cultura, porque aí qualquer coisa que nos venha de fora, em vez de ser influência que nos esmaga, que nos descaracteriza, que nos corrrompe, passa a ser uma incorporação que nos enriquece».

Ariano Suassuna



#salveoqueijoartesanal

As regiões assinaladas ilustram apenas parcialmente a complexidade e a diversidade da produção artesanal de queijos no Brasil. Mapa produzido para divulgação na internet e mídias digitais. Entre em contato para versões impressas e publicações. © Slow Food Brasil 2018.

Você conhece algum queijo artesanal que não está nesse mapa? Então compartilhe! [queijoartesanal@slowfoodbrasil.com](mailto:queijoartesanal@slowfoodbrasil.com)



Figura 1 – Mapa queijos artesanais. Fonte: (FOOD, 2018).

O bioma da Mata Atlântica ocupa grande parte do litoral brasileiro e se estende completamente nos estados do Rio de Janeiro, Espírito Santo e Santa Catarina, e parcialmente nos estados do Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia, Minas Gerais, Goiás, Mato Grosso do Sul, São Paulo, Paraná e Rio Grande do Sul. Aqui 15 queijos possuem destaque que são o Queijo Cabacinha, o Requeijão, o Queijo Minas Serro, o Queijo Minas Campo das Vertentes, o Parmesão da Mantiqueira, o Parmesão de Alagoa, o Queijo Minas Canastra, o Queijo Porungo, o Requeijão de Prato, o Kochkäse, o Queijo Diamante, o Queijinho, o Queijo Serrano, o Queijo do Contestado e o Queijo Colonial, que assim como o Coalho na região nordeste do Brasil possui diversas variedades de região para região no sul do País.

O bioma dos Pampas se encontra de modo exclusivo no estado do Rio Grande do Sul se estendendo em todo o Uruguai, centro-leste da Argentina e extremo sudeste do Paraguai. De acordo com Food (2018) não há nenhum queijo artesanal de destaque nesta região.

É possível observar que há uma grande variedade de tipos de queijos artesanais espalhados por todas as regiões do Brasil e toda essa variedade se dá por diversos fatores como o tipo de leite a ser utilizado (se é de vaca, cabra, ovelha ou búfala), o bioma, onde as condições climáticas específicas e presentes em cada um proporciona condições distintas para a proliferação de microrganismos vivos responsáveis pelo processo de maturação do queijo, a adição ou não de coalho, o tempo de cura, e o processo manual de se produzir o queijo, são peças chave e responsáveis pela grande variedade de características como o sabor, textura, cor, odor, fungos, dentre outros agradando assim um público bastante amplo e diversificado.

Com o intuito de divulgar e mostrar o quão grande esse mercado de queijos artesanais no Brasil têm-se tornado, em 2017 a Associação de Comerciantes de Queijos Artesanais Brasileiros - COMERQUEIJO, que é um órgão constituído por comerciantes de todo o País que se dedicam ao comércio de queijo artesanal idealizaram o Prêmio Queijo Brasil, que é um concurso nacional que visa fortalecer e divulgar os queijos artesanais brasileiros expandindo assim o mercado consumidor desses produtos. A primeira edição deste concurso ocorreu em 2014 com 136 queijos inscritos. Já em 2016 o número de inscrições subiu para 234 e em 2017 na terceira edição o número de inscrições atingiu a marca de 363 queijos, evidenciando a força desse mercado. Em 2018, esse número de inscrições foi ainda maior, chegando a 479 queijos inscritos, mas o grande recorde inscrições foi no ano de 2019 onde cerca de 718 queijos foram inscritos mostrando o quanto este produto vem crescendo em representatividade e aumentando a presença nas casas dos brasileiros (COMERQUEIJO, 2019).

### 2.1.1 Características dos Tipos de Queijos Artesanais

Com intuito de separar os tipos de queijos artesanais em categorias, foi utilizado a categorização feita pelo Concurso Prêmio Queijo Brasil e mais duas categorias que não estava presentes neste concurso que são a categoria Artesanal Paulista e Outros.

Para que um queijo possa participar da premiação do Concurso Prêmio Queijo Brasil, é necessário que seus produtores e/ou representantes façam a inscrição do queijo alguma das categorias: Minas Artesanal, Alagoa, Coalho, Colonial, Manteiga, Mantiqueira de Minas, Requeijão, Serrano, Cabacinha, Contestado, Kochkäse, Marajó, Nicola, Porongo, Queijinho, Maturado que serão melhor detalhadas a seguir (COMERQUEIJO, 2019).

A categoria Artesanal Paulista, engloba queijos artesanais de destaque pertencentes ao estado de São Paulo. A categoria Outros foi criada devido a algumas palavras-chave referentes aos nomes de queijos artesanais não pertencerem a nenhuma das categorias descritas no Concurso Prêmio Queijo Brasil.

#### 2.1.1.1 Minas Artesanal

O Queijo Minas Artesanal (QMA) é um produto tipicamente mineiro e sua principal característica está relacionada com os atributos sensoriais que este produto possui referente à presença de microrganismos originais constituintes da parte microbiota natural do leite que é responsável pela promoção de diferentes características como sabor, textura e aroma (DORES, 2013). Estes microrganismos presentes nos queijos variam de acordo com clima, alimentação do rebanho e a forma de processamento do produto fazendo com que o QMA do Serro seja diferente do QMA da Serra da Canastra que por sua vez é diferente do QMA de Araxá que é distinto do QMA do Cerrado (MONTEIRO, 2018).

Nestas regiões encontram-se diversos queijos conhecidos mundialmente devido a prêmios conquistados em concursos internacionais sendo que em 2019, 50 produtores de queijos artesanais de Minas Gerais conquistaram medalhas no 4º Concurso “*Mondial du Fromage et des Produits Laitiers*” realizado na França, mostrando assim o quão esses produtos são diferenciados e importantes para o povo mineiro (MELO et al., 2019).

#### 2.1.1.2 Artesanal Paulista

A categoria Artesanal Paulista foi criada para este trabalho por São Paulo possuir diversos queijos artesanais de destaque. Dessa maneira, grande parte dos queijos mais conhecidos pertencem ao denominado Caminho do Queijo Artesanal Paulista que foi criado em 2017, como forma de valorizar a produção queijeira de São Paulo mostrando a todos que essa região é capaz de produzir queijos tão bons quanto aos de outros estados e países (PAULISTA, 2021).

Atualmente, 12 das queijarias artesanais paulistas mais conhecidas fazem parte

deste caminho que são : Fazenda Atalaia (Amparo), Fazenda Santa Helena (Sete Barras), Fazenda Santa Luzia (Itapetininga), Laticínio Artesanal Montezuma (São João da Boa Vista), Laticínio Artesanal (Pardinho), Leitaria Santa Paula (São José do Rio Pardo), Pardinho Artesanal (Pardinho), Pé do Morro (Cabreúva), Queijaria Belafazenda (Bofete), Queijaria Rima (Porto Feliz), Cabanha Mulekinha (Ibiúna), Lano-Alto (Catuçaba) e Q.JO Martina (Boitúva) (PAULISTA, 2021). No manifesto criado pelos participantes e criadores do Caminho do Queijo Artesanal Paulista, é possível encontrar as especificações e regras a serem seguidas para ser considerado um queijo artesanal paulista.

#### 2.1.1.3 Alagoa

O queijo artesanal de Alagoa possui o sabor da natureza devido sua produção ser realizada em áreas preservadas da Serra da Mantiqueira no sul de Minas Gerais (APROALAGOA, 2021). Sua história se inicia no começo do século XX sendo fabricado em pequenas propriedades e sítios pelas mãos dos próprios familiares. A tecnologia foi herdada de uma família de italianos que veio morar em Alagoa e ao perceberem que o clima frio da região, era parecido com o norte da Itália, resolveram experimentar uma receita de família na confecção do tradicional queijo do norte da península (APROALAGOA, 2021). O fato é que ao longo dos anos este queijo vem ganhando destaque não só nacional mas também internacional, com diversos prêmios em concursos mostrando a qualidade deste produto (APROALAGOA, 2021).

#### 2.1.1.4 Coalho

O queijo Coalho é bastante difundido no Brasil e possui uma longa história e estima-se que o mesmo é reconhecido como tal há mais de 150 anos (CAVALCANTE, 2007), e foi criado após longas jornadas de viajantes que acondicionavam leite ao matulão (bolsa feita do estômago de animais), que se coagulava formando a massa chamada de Queijo de Coalho (CHAVES et al., 2021). O queijo coalho artesanal é produzido em todo o Brasil. Apesar de controversa, a origem desse derivado lácteo é atribuída à região Nordeste, onde ainda é fabricado tradicionalmente utilizando o leite cru (FOOD, 2018). De modo geral, no Norte e Nordeste do Brasil a forma como é consumido é bastante diferente das demais regiões brasileiras, sendo encontrado no café da manhã, como tira gosto em mesas de bares e em comidas típicas, como é o caso do Baião de Dois (MANDACARU, 2021). Já nas regiões Sul e Sudeste é possível encontrar em padarias, atacados e supermercados na forma de espetinhos simples, com ou sem orégano em pequenos pedaços ou em peças de até 3,5 kg (MANDACARU, 2021).

#### 2.1.1.5 Colonial

O queijo Colonial é fabricado na região serrana do estado do Rio Grande do Sul sendo também fabricado em Santa Catarina, e é um dos mais famosos ícones da gastronomia gaúcha. Com um sabor láctico e medianamente picante com uma coloração amarela pálida é bastante consumido acompanhado de vinhos tintos ou utilizado em pizzas, massas e gratinados (DELHAYE, 2016).

#### 2.1.1.6 Manteiga

O queijo de Manteiga é um queijo tipicamente nordestino. O Queijo de Manteiga é obtido mediante coagulação do leite com emprego de ácidos orgânicos de grau alimentício, cuja massa é submetida à dessoragem, lavagem e fusão, com acréscimo exclusivamente de manteiga de garrafa ou manteiga da terra ou manteiga do sertão (BRASIL, 2001).

#### 2.1.1.7 Mantiqueira de Minas

Mesmo utilizando o leite cru no processo de fabricação, há uma diferença entre o Queijo Minas Artesanal e o Queijo Artesanal Mantiqueira de Minas. No primeiro, utiliza-se o pingo<sup>1</sup> e um modo de fazer diferente do Queijo Artesanal Mantiqueira de Minas. Este utiliza fermentos ao invés do pingo e há uma elevação de temperatura durante a sua produção. A região produtora do Queijo Artesanal Mantiqueira de Minas engloba 9 cidades que são: Aiuruoca, Baependi, Bocaina de Minas, Carvalhos, Itamonte, Liberdade, Itanhandu, Passa Quatro e Pouso Alto (RODOLFHO, 2020).

#### 2.1.1.8 Requeijão

A categoria Região possui duas palavras-chave referentes aos queijos denominados: Requeijão do Sertão e Requeijão Moreno. O primeiro é caracterizado como um queijo defumado com alto teor de gordura sendo encontrado no sertão baiano, Sergipe e na porção Sul nas proximidades do norte de Minas Gerais (SOUZA, 2016). O Requeijão Moreno é encontrado no norte de Minas Gerais. De acordo com Franco (2018), as principais formas de consumo deste produto são: acompanhado por uma xícara de café, por algum doce em calda ou derretido em uma frigideira e finalizado com açúcar ou melaço de cana.

#### 2.1.1.9 Serrano

Os queijos Colonial e Serrano são típicos do Rio Grande do Sul e o fator que os difere está na área de abrangência da produção. Enquanto o Colonial é encontrado em

---

<sup>1</sup> O “pingo” de acordo com (IPHAN, 2014) é um fermento natural que foi desenvolvido ao longo dos tempos, a partir do soro drenado do próprio queijo. É responsável por fornecer aos queijos características microbiológicas específicas, condicionadas pelo tipo de solo, clima e vegetação de cada lugar.

todo o Estado, o Serrano é restrito a dois locais específicos que são os Campos de Cima da Serra e uma parte de Santa Catarina (MANDARINO, 2018).

Esse produto é caracterizado por uma coloração amarelada e uniforme, com odor marcante e agradável apresenta um sabor ligeiramente ácido, amanteigado e picante variando conforme o período de maturação (WAGNER, 2018; COUTINHO et al., 2018).

#### 2.1.1.10 Cabacinha

O queijo Cabacinha, também conhecido como *Caccio Cavallo* na Itália, recebe esse nome devido ao seu formato após ser amarrado e pendurado para secar. Produzido no Vale do Jequitinhonha, é muito famoso na região e é muito utilizado em pizzas, sanduíches e recheados, da mesma forma que o Provolone. De acordo com Queijos (2020), este queijo combina com azeite de oliva, saladas, tomate e orégano podendo ser consumido também com vinhos de taninos marcantes como o *Cabernet Sauvignon* ou de uvas brancas secas.

#### 2.1.1.11 Contestado

O queijo Contestado foi um dos premiados com medalha de prata no 5º Prêmio Queijo Brasil e é produzido na cidade de Caçador - Santa Catarina. Não se obteve muitas informações sobre este queijo, mas sabe-se que em 2017 a Prefeitura Municipal de Caçador buscava informações para criação da Lei do Queijo Artesanal do Contestado devido às suas características peculiares diferentes do Queijo Artesanal Serrano, e como forma de valorizar os produtores de queijo de Caçador tirando-os da informalidade (CAÇADOR, 2017).

#### 2.1.1.12 Kochkäse

O queijo Kochkäse é historicamente produzido nas áreas de colonização alemã em Santa Catarina, principalmente no Vale do Itajaí (ITTNER, 2020). Possui sabor e odor característicos, com uma consistência mole, cor amarelada. Tais características necessitam ser respeitadas de acordo com a regulamentação do mesmo, que tira da ilegalidade os produtores favorecendo assim mais de 6 mil famílias catarinenses (ITTNER, 2020).

#### 2.1.1.13 Marajó

O queijo da Ilha do Marajó no Estado do Pará é produzido de duas formas, que são manteiga e creme e não passam pelo processo de maturação. A técnica de produção artesanal do Queijo do Marajó é única no mundo, pelo fato de utilizar somente o leite de búfala. Com isso, este queijo já conquistou diversos prêmios com as medalhas de Bronze em 2017, Super Ouro em 2018 e Ouro em 2019 no Prêmio Queijo do Brasil e Prata na quarta edição do “*Mondial du Fromage et des Produits Laitiers*”, na França (LAVOURA, 2020).

#### 2.1.1.14 Nicola

O queijo Nicola, queijo possui textura semelhante à muçarela derretida e lembra também o queijo Cabacinha, também conhecido como *caccio cavallo* na Itália pelo processo de secagem pendurado (TORRES, 2017). Seu tempo de cura é menor e depende do objetivo do produtor. Este nome, faz referência à região do Pantanal, nas áreas de Urucum e Jacadigo, homenageando o distrito de Nhecolândia (TORRES, 2017). A receita do Queijo Nicola foi elaborada por peões de comitiva durante a ocupação do Pantanal, local este que foi precursor na produção do queijo (TORRES, 2017).

#### 2.1.1.15 Porongo

O queijo Porongo recebe esse nome devido ao seu formato parecer com um fruto chamado porongo ou cabaça, que é um fruto da família das cucurbitáceas, natural da África e da Índia (CASTRO, 2013). Em outras regiões do Brasil é também chamado de queijo cabacinha que já foi descrito anteriormente.

#### 2.1.1.16 Queijinho

O Queijinho se parece muito com a ricota, mas a ricota é produzida a partir do soro do leite, enquanto o Queijinho também conhecido como Queijinho Branco é fabricado a partir da coagulação do leite, ou seja, a mesma base utilizada para a produção do queijo “Kochkäse” (MARESCH, 2020). Geralmente é consumido com pão, tortas doces ou salgadas, sendo muito popular por sua característica acidulada de sabor (MARESCH, 2020).

#### 2.1.1.17 Maturado

A categoria Maturado foi criada pelo fato de alguns queijos presentes na lista de queijos premiados do concurso “Prêmio Queijo Brasil” (Apêndice A) só apresentarem a informação que são maturados.

#### 2.1.1.18 Outros

A categoria outros, aborda queijos e palavras-chave do universo dos queijos que não pertencem a nenhuma das categorias descritas anteriormente, sendo utilizados para coletar mais dados que pertençam ao contexto dos queijos artesanais.

### 2.1.2 Efeitos da Pandemia no Consumo

Um fato que recentemente vem sendo muito comentado e compartilhado em todos os meios de comunicação inclusive nas redes sociais é a pandemia do novo coronavírus (COVID-19), que teve início em Wuhan, capital da província de Hubei na China, em

novembro de 2019. Relatado oficialmente no dia 31 de dezembro de 2019, rapidamente se alastrou por todo o mundo sendo reconhecido como pandemia global no dia 11 de março de 2020. Devido ao alto poder de contágio da doença, diversas medidas de isolamento social e quarentena foram impostas em diversos países do mundo com o intuito de desacelerar o crescimento da doença enquanto não se desenvolvia a vacina para a mesma.

Diversas mudanças ocorreram na rotina das pessoas, e segundo Siqueira e Rocha (2020), durante a pandemia da COVID-19, o queijo foi o derivado lácteo mais consumido pelos respondentes da pesquisa. De acordo com a pesquisa de mercado da Nielsen (MILKPOINT, 2020), entre maio de 2019 e o mesmo período em 2020 houve aumento de 29,1% no consumo de queijos, 17,6% no leite UHT e de 16,4% no leite em pó.

Ao contrário de outros segmentos, como pôde ser observado por estas pesquisas, o setor de leite e derivados durante a pandemia obteve resultados bastante expressivos. Com o grande número de pessoas em casa em isolamento social, houve no Brasil segundo dados do *Google Trends*, um aumento considerável em 2020 no interesse dos usuários pelos termos de busca *lanches*, *receitas*, e *delivery*, que pode ser visto na Figura 2 e como o queijo é um ingrediente chave em muitos pratos e lanches entregues por *delivery*, acredita-se que esses dados reforçam ainda mais a afirmação de aumento no consumo deste produto.

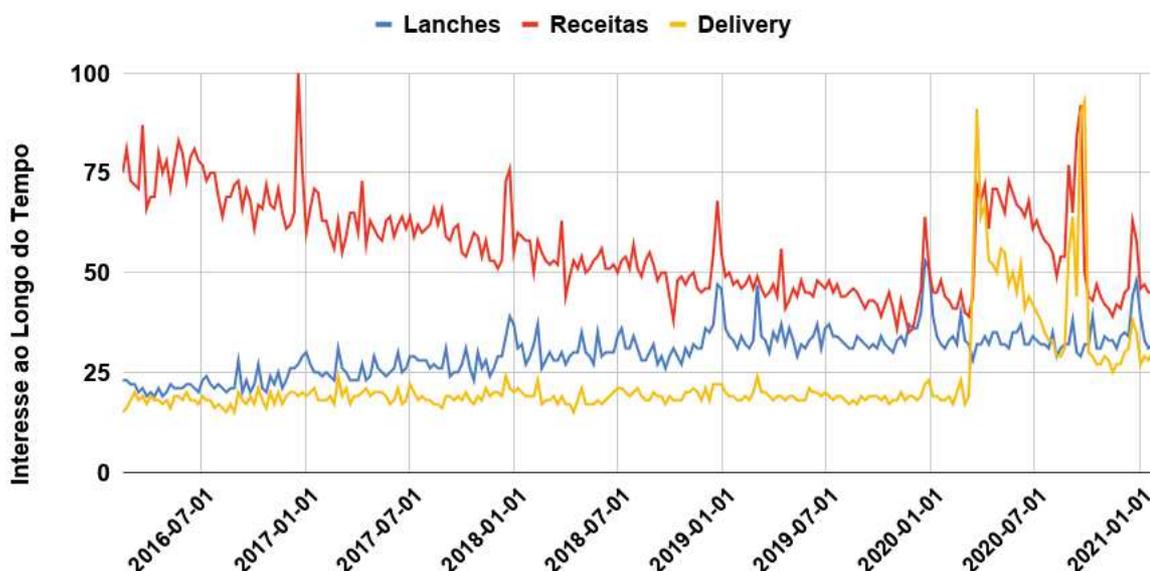


Figura 2 – Interesse pelos termos de busca *lanches*, *receitas* e *delivery* nos últimos 5 anos no Brasil. **Fonte:** (TRENDS, 2021).

Como pode ser visto, a mudança temporária na rotina das pessoas, trouxe alterações consideráveis nos hábitos alimentares. A tendência que antes se tinha na busca de uma alimentação saudável e rica em nutrientes mudou e segundo Siqueira et al. (2020a), o cenário brasileiro nas primeiras semanas de pandemia foi marcado pelo aumento na procura

de alimentos considerados indulgentes (aqueles capazes de fornecer sensação de prazer). Com todo mundo dentro de casa e impossibilitados de realizar as atividades de costume, estes alimentos tornaram-se um novo meio de aliviar o *stress* causado pelo isolamento social.

## 2.2 TENDÊNCIAS DE CONSUMO DE ALIMENTOS

Com consumidores cada vez mais exigentes, saber como agradá-los e desenvolver melhores produtos e serviços é de extrema importância para minimizar as chances de sucesso da concorrência. Entender bem, qual o rumo que o mercado está seguindo nos dias de hoje é de fundamental importância para atender as necessidades do consumidor. Definido como a “propensão dos indivíduos em modificar hábitos já estabelecidos” (BARBOSA et al., 2020), o termo tendência busca entender quais são as preferências dos consumidores, identificando características e hábitos de consumo que buscam relação a algum determinado produto e/ou serviço.

De acordo com o resultado das análises comparativas dos estudos sobre tendência presente no trabalho de Barbosa et al. (2020), foram mapeadas 5 grandes macrotendências da alimentação que são: **Sensorialidade e Prazer**, **Saudabilidade e Bem-estar**, **Conveniência e Praticidade**, **Confiabilidade e Qualidade**, **Sustentabilidade e Ética**. Em Zacarchenco et al. (2020), mais uma categoria foi pensada e denominada **Premiumização<sup>2</sup> e Sensorialidade**.

Ainda segundo Barbosa et al. (2020), com relação à **Sensorialidade e Prazer**, as tendências estão relacionadas com a melhoria de aspectos na vida das pessoas como a educação, o acesso à informação, o aumento de renda e outros fatores. Assim, os consumidores valorizam novas experiências alimentares buscando cada vez mais, produtos novos que sejam saborosos, saudáveis e exóticos. A categoria **Saudabilidade e Bem-estar** refere-se a fatores como o envelhecimento de uma determinada população, onde uma das maiores preocupações está relacionada à busca por alimentos funcionais, produtos para dietas e controle do peso como meio de se obter um estilo de vida mais saudável. Com a vida agitada nos centros urbanos, as tendências relacionadas à **Conveniência e Praticidade** se referem a indivíduos que buscam refeições que sejam prontas e/ou semi-prontas para o consumo, com embalagens práticas de fácil abertura, rápido preparo destacando-se produtos em pequenas porções. Já as tendências relacionadas à **Confiabilidade e Qualidade** agrupam um grupo de consumidores preocupados com segurança e qualidade. São consumidores que gostam de saber a origem de onde vem os alimentos que vão consumir ou que consomem, valorizam selos de qualidade, e para conquistá-los as marcas necessitam passar confiança e credibilidade através destes requisitos. Com relação à

---

<sup>2</sup> É um neologismo e refere-se à busca de produtos mais sofisticados, luxuosos e com maior valor agregado.

**Sustentabilidade e Ética** verifica-se um perfil de consumidores preocupados com o meio ambiente e pontos como embalagens recicláveis e baixos níveis de impactos ambientais são bem vistos. Com relação à **Premiumização e Sensorialidade** que seria a sexta categoria de tendências de consumo de alimentos, segundo Zacarchenco et al. (2020), com relação a **Premiumização** os consumidores buscam produtos diferenciados e sofisticados com um maior valor agregado, valorizando produtos regionais e artesanais com certa personalização e com respeito à **Sensorialidade**, características parecidas presentes nos estudos de (BARBOSA et al., 2020) foram observados onde o fator mais importante está relacionado a experiências sensoriais sofisticadas e inovadoras com um escape do cotidiano através de comidas exóticas.

### 2.3 BUSINESS INTELLIGENCE

Com o intuito de apoiar e facilitar a maneira de se gerenciar uma empresa, o *Business Intelligence* (BI), Inteligência de Negócios ou Inteligência Empresarial atualmente, vem sendo bastante utilizado por empresas como um importante meio de aquisição de informações dos setores que compõem uma determinada organização. Um dos processos mais cruciais em um ambiente empresarial é o de tomada de decisões, e com a grande quantidade de dados que são gerados minuto a minuto, as organizações têm gastado bastante energia na construção de metodologias capazes de transformar dados em informações a fim de possibilitar uma visão mais ampla do negócio e a construção de estratégias mais assertivas e direcionadas.

Angeloni e Reis (2006) definem o *Business Intelligence* como um conjunto de metodologias de gestão que são implementadas por meio de um *software*, com a principal função de gerar ganhos em processos de tomada de decisão no gerenciamento da administração de organizações, integrando em um único lugar informações relevantes para a decisão a ser tomada. Para Raisinghani (2004), BI é um termo “guarda-chuva” que aborda arquiteturas, ferramentas, bancos de dados, aplicações e metodologias. Com isso, fica ainda mais visível que a implementação do BI nas organizações, torna a aquisição de conhecimento mais assertiva através da manipulação de dados, transformando estes em informação, e informação em ação.

Nos últimos anos, o BI evoluiu bastante com o objetivo de tornar o seu desempenho ainda melhor. De acordo com Tableau (2021) a mineração de dados faz uso de banco de dados, estatística e técnicas de aprendizado de máquina para identificar tendências em grandes volumes de dados. Atividades como a geração de relatórios, são responsáveis por exibir as análises de dados realizadas para os responsáveis de determinada área com o intuito de facilitar o entendimento de conclusões e tomada de decisões. *Benchmarking* e métricas de desempenho têm a função de comparar dados históricos tornando simples o acompanhamento de metas. A análise descritiva utiliza a análise de dados para descobrir

o que aconteceu em determinado momento, a fim de responder perguntas específicas com respostas presentes nos conjuntos de dados. A análise estatística tem como objetivo explorar os dados com maior profundidade fazendo o uso de técnicas de visualização dos dados a fim de transformar números em gráficos facilitando a obtenção da informação de forma direta.

A utilização de BI tem sido bastante explorada e diversos setores obtiveram sucesso com tais aplicações. De acordo com Abel (2018), na Copa do Mundo de Futebol de 2014, a seleção alemã utilizou o BI como uma ferramenta de auxílio para verificação do desempenho do time ao longo da competição. Analisando dados como o número de passes, velocidade em campo, finalizações, defesas e penalidades dos jogadores foi possível identificar quais atletas possuíam os melhores rendimentos, tornando possível sempre escalar o melhor time titular para cada partida. Assim, a aplicação de BI foi um dos fatores determinantes para a conquista da Copa do Mundo de 2014 pela seleção alemã, servindo de inspiração para outros clubes em todo mundo. Outro caso de sucesso da aplicação de BI, ainda de acordo com Abel (2018), é o caso da Fiat Automóveis. O Uno é um dos principais carros da Fiat e o desafio que tinham para 2011 era o de deixar o carro com a “cara do consumidor”. Para isso, a Fiat realizou diversas pesquisas de mercado com o intuito de levantar dados sobre seus seguidores nas redes sociais. A partir daí, com a utilização de BI para a realização da análise dos dados coletados, a empresa conseguiu absorver as expectativas dos consumidores e com base nessas informações, desenvolveram o novo Uno, que recebeu diversos prêmios sendo o Carro do Ano de 2011 da revista Auto Esporte.

## 2.4 INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial, também conhecida pela sigla IA, é uma das áreas da ciência da computação que tem por finalidade tentar descrever por meio de algoritmos uma simulação da inteligência humana na hora da tomada de decisões. Dessa forma, diversos autores buscaram de alguma maneira definir tal conceito. De acordo com John McCarthy, considerado o pai deste conceito, “A Inteligência Artificial é a ciência e a engenharia de criar máquinas inteligentes, especialmente programas de computador inteligentes”. Outra definição bastante interessante é descrita por Luger (2004) e diz que a inteligência artificial “pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento inteligente”. Barr e Feigenbaum (1981) diz ainda que “a IA é parte da ciência da computação voltada para o desenvolvimento de sistemas de computadores inteligentes, *i.e.* sistemas que exibem características, as quais nós associamos com a inteligência no comportamento humano - ex. compreensão da linguagem, aprendizado, raciocínio, resolução de problemas etc.”.

Sendo assim, a IA vem se desenvolvendo ao longo dos anos com o objetivo de desenvolver tarefas que atualmente somente humanos desenvolvem com excelência. Nos

últimos 5 anos, segundo dados do Google Trends (2021), é possível verificar que houve um grande aumento do interesse de usuários pelo termo de pesquisa Inteligência Artificial, o que pode ser visto na Figura 3.

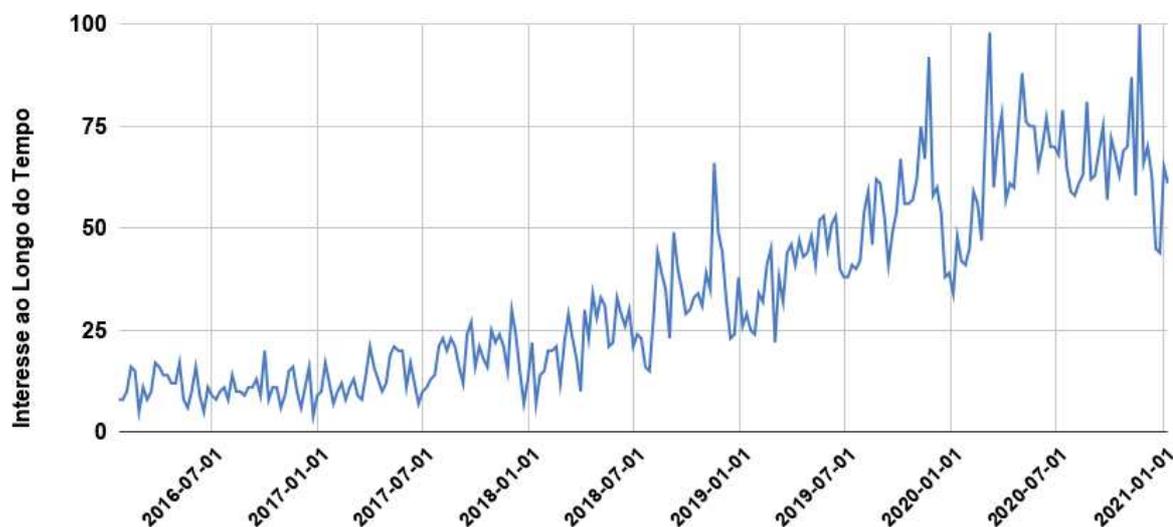


Figura 3 – Interesse pelo termo de pesquisa Inteligência Artificial no Brasil entre 2016 e 2021. **Fonte:** Google Trends (2021).

A IA está presente nos dias de hoje em diversas aplicações que vão desde o assistente virtual até as aplicações de visão computacional para reconhecimento de objetos e fala. Está presente também em sistemas de recomendação baseados no comportamento do usuário, detecção de fraudes e na análise de sentimento presente em vídeos, imagens ou textos, que será o nosso foco neste trabalho. O grande número de aplicações que é possível de ser feito com técnicas de IA podem ser divididas em modelos com as expertises de simular o aprendizado e raciocínio humano. No aprendizado de máquina (do inglês *Machine Learning*) tem-se algoritmos capazes de aprender de forma automática sem ser necessário programar tais formas de aprendizado, e na aprendizagem profunda (do inglês *Deep Learning*), que é uma subárea do aprendizado de máquina, as redes neurais artificiais se adequam na medida em que vai se aprendendo com uma grande quantidade de dados. A Figura 4 exhibe um esquema visual mostrando o relacionamento entre as áreas de estudo.

#### 2.4.1 Aprendizado de Máquina

Dentro da Inteligência Artificial, uma área de estudo que detém um destaque bastante expressivo é a de aprendizado de máquina. Segundo Faceli et al. (2011), este método visa desenvolver programas com a capacidade de aprender e executar tarefas de acordo com sua própria experiência. Dessa forma, utilizando dados que representam um determinado tipo de experiência, o algoritmo aprende e se torna capaz de realizar uma determinada ação com base no aprendizado que obteve.

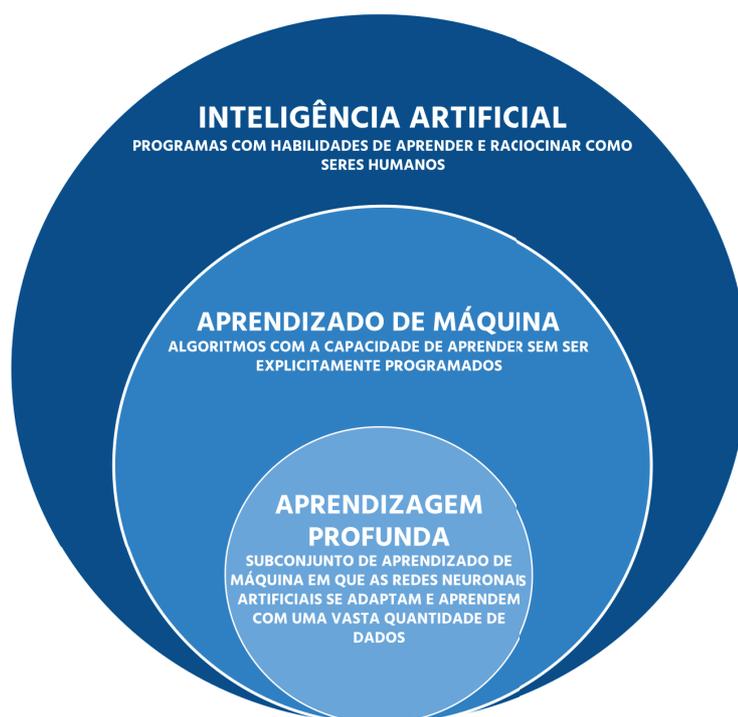


Figura 4 – Hierarquia das áreas de estudo. **Fonte:** Baseado em Santana (2018).

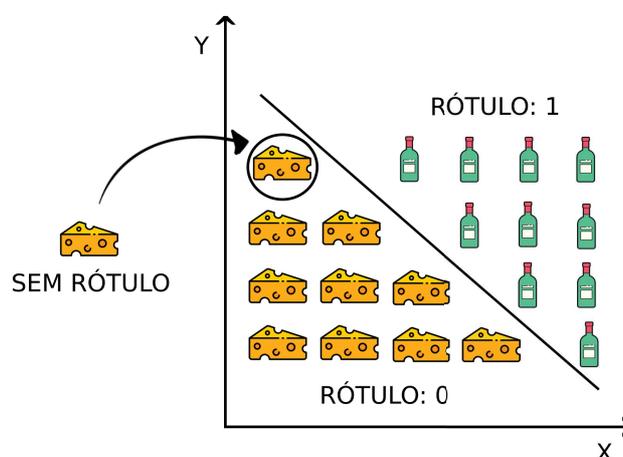
A área de aprendizado de máquina pode ser dividida em dois grupos extremamente importantes que são os de aprendizagem supervisionada e não-supervisionada.

#### 2.4.1.1 Aprendizado Supervisionado

Desde pequeno, o ser humano passa pelo processo de aprendizagem e logo nos primeiros instantes de vida precisa aprender a sobreviver e dessa forma, atividades como respirar, chorar, se alimentar são essenciais para ele. Com o passar do tempo, com a **supervisão** de seus pais, a criança aprende a falar, andar, brincar, comer até de fato estar apta a aprender por conta própria diversas outras coisas. Da mesma maneira, no aprendizado supervisionado, como o próprio nome diz, é necessário fornecer ao algoritmo um conjunto de dados rotulados referentes à aplicação a qual deseja-se realizar para que logo em seguida a máquina consiga aprender e realizar as tarefas por conta própria.

Dessa maneira, de acordo com Maimon e Rokach (2005), os métodos de aprendizado supervisionado são aqueles que tentam descobrir a relação existente entre os atributos de entrada (conhecidos também como variáveis independentes) e uma classe de destino (também chamada de variável dependente). Existem dois modelos de aprendizagem supervisionada bastante conhecidos, o primeiro se refere à classificação onde é realizada a predição para uma determinada classe, assumindo valores discretos e a segunda referente aos modelos de regressão onde a variável alvo é contínua. Pode-se pensar no seguinte exemplo abaixo:

**Exemplo:** Seu chefe lhe dá um desafio e você precisa encontrar padrões em um conjunto com 20.000 imagens sendo 10.000 imagens de queijos e 10.000 de garrafas de vinho. Cada imagem possui um rótulo onde 0 se refere às imagens de queijo e 1 às garrafas de vinho. Dessa maneira, agora tem-se que o objetivo é o de utilizar a técnica de aprendizado supervisionado para encontrar padrões nas imagens dividindo o conjunto de dados em duas metades. Dessa forma, quando o algoritmo receber uma nova imagem digamos que a de um queijo, sem qualquer rótulo, ele precisa classificá-la como uma imagem de queijo de forma correta mesmo sem a presença de nenhum rótulo, baseando-se no aprendizado que obteve no treinamento. A Figura 5, exibe o processo do aprendizado supervisionado.

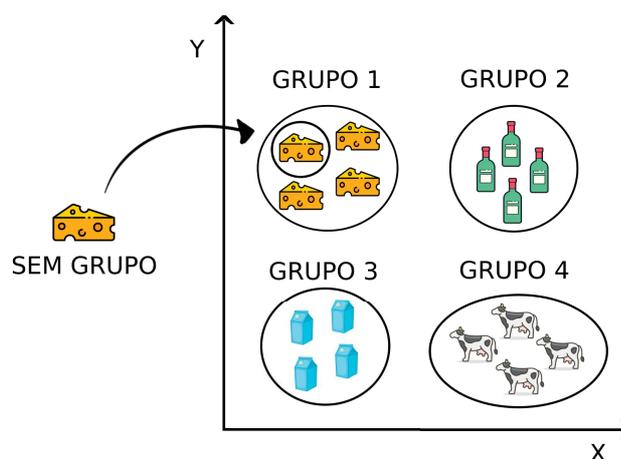


**Figura 5** – Exemplo de aprendizado supervisionado. **Fonte:** Baseado em Sharma (2018).

#### 2.4.1.2 Aprendizado Não-supervisionado

Diferentemente dos métodos de aprendizado supervisionado, neste caso, não há nenhuma informação adicional ao modelo, ou seja, o conjunto de dados não possui rótulos para o aprendizado do mesmo. Dessa forma, a ideia desses modelos é o de realizar agrupamentos com base em características presentes no conjunto de dados, colocando no mesmo grupo os mais parecidos. Segundo Duda et al. (2001), para cada grupo, ou *cluster*, utiliza-se um rótulo, tornando possível indicar a qual grupo cada registro pertence. Vejamos o seguinte exemplo abaixo:

**Exemplo:** Mais uma vez seu chefe lhe propõe um desafio só que agora lhe apresenta um conjunto com 20.000 imagens sendo 5.000 de queijos, 5.000 de garrafas de vinho, 5.000 de caixas de leite e 5.000 de vacas. Porém, ele não lhe fornece nenhuma informação adicional como descrito no aprendizado supervisionado que são os rótulos. Dessa forma, é necessário deixar o algoritmo aprender por conta própria verificando quais são as características de cada imagem agrupando-as em grupos (*clusters*) da melhor maneira que conseguir. A Figura 6, mostra como seria esse tipo de abordagem.



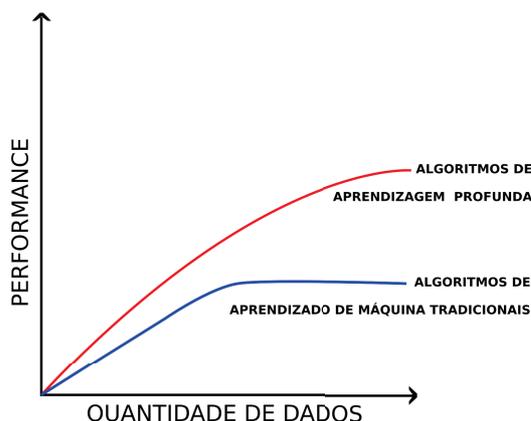
**Figura 6** – Exemplo de aprendizado não-supervisionado. **Fonte:** Baseado em Sharma (2018).

#### 2.4.2 Aprendizagem Profunda

A aprendizagem profunda é considerada como uma tecnologia ininterrupta de aprendizado que possui um elevado desempenho na resolução de problemas (COPELAND, 2016). Este termo é usado para denotar o problema de treinar redes neurais artificiais que realizam o aprendizado de características de forma hierárquica, de tal forma que características nos níveis mais altos da hierarquia sejam formadas pela combinação de características de mais baixo nível. Essa abordagem busca imitar o cérebro humano o processamento de dados do cérebro humano na criação de padrões para posteriores tomadas de decisão. A grande diferença entre os demais métodos apresentados é a necessidade de uma grande quantidade de dados, o que para determinadas aplicações pode ser bastante custoso de se obter, mas que, quando se possui, os resultados relacionados quanto ao desempenho são mais efetivos. A Figura 7 mostra a performance dos algoritmos de aprendizado de máquina tradicionais e os algoritmos de aprendizagem profunda em relação a quantidade de dados.

### 2.5 MINERAÇÃO DE DADOS

É fato que a popularização da internet impactou e muito na quantidade de dados gerados a todo momento. Com o alto número de pessoas conectadas na rede por meio de diversos dispositivos como os celulares, *tablets*, televisores, computadores, dentre outros, gerenciar essa imensidão de dados a fim de transformá-los em informação, tornou-se um grande desafio. De acordo com MJV (2020), as redes sociais do Facebook e Twitter geram por minuto, cerca de 510.000 comentários e 347.000 *tweets* respectivamente, mostrando o quão esses meios são utilizados pelas pessoas e o tanto de informações que podem estar presentes nessa imensidão de dados.



**Figura 7** – Comparação entre algoritmos de aprendizado de máquina tradicionais e de aprendizagem profunda. **Fonte:** Baseado em Sharma (2018).

A técnica de mineração de dados (do inglês *Data Mining*), segundo Avelar et al. (2017), se refere ao processo de extração de conhecimento a partir de grandes volumes de dados, que possibilitam a descoberta de padrões e possíveis relações ocultas que auxiliam na construção de regras e correlação de dados tornando a tomada de decisão mais assertiva de forma rápida e confiável. Avelar et al. (2017) cita referências de trabalhos que mostram diversas tarefas que podem ser realizadas com o uso da mineração de dados (DIAS, 2002; HAIR et al., 2003; GALVÃO; MARIN, 2009), sendo que as de maior uso são:

- Classificação: responsável pela construção de um modelo que possa ser aplicado a uma base de dados não classificados, possibilitando assim categorizá-los em classes a fim de descobrir um relacionamento existente entre um atributo real com um atributo previsto;
- Estimativa ou regressão: é responsável por definir um valor para alguma variável contínua desconhecida, um exemplo seria o de estimar o tempo de vida de um cliente;
- Associação: é utilizado para determinar quais itens tendem a ser adquiridos em uma mesma transação;
- Segmentação: se refere a divisão de uma determinada população heterogênea em subgrupos ou grupos mais homogêneos;
- Sumarização: desenvolve métodos para a encontrar uma descrição compacta para um subconjunto de dados;
- Séries temporais: parecido com a técnica de regressão, porém utiliza medidas baseadas no tempo, verificando relações que podem existir em determinados períodos.

Minerar dados textuais é uma tarefa que aborda diversas áreas da informática e segundo Aydano (2010) áreas como aprendizado de máquina, recuperação de informação,

estatística e linguagem computacional são utilizadas a fim de tornar possível a transformação da linguagem humana textual em linguagem de máquina que o computador consegue entender.

### 2.5.1 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural, é uma das principais técnicas utilizadas para extração de informações de dados textuais. Segundo Covington (1994), Russell e Norvig (1995), o PLN consiste no desenvolvimento de modelos computacionais para a realização de atividades que dependem de informações expressas em alguma língua natural (por exemplo, tradução e interpretação de textos, busca de informações em documentos e interface homem-máquina).

O fato é que este processamento não é tão simples e segundo Finzi (2019) a linguagem natural (tanto escrita, quanto falada) possui diversas regras e ambiguidades, que dependem inclusive do idioma. O formalismo empregado no texto também traz diversos impactos neste processo, só o fato de analisar um texto bem formatado no padrão culto da língua portuguesa ou textos informais presentes em redes sociais causam um grande impacto nos resultados. Com o uso de um linguajar em sua grande maioria informal, ao analisar dados das redes sociais, é possível verificar a presença de palavras abreviadas, palavras escritas de forma errada, uso de *emojis* (carinhas que exibem uma determinada expressão), ironia, sarcasmo, duplo sentido que quando presentes em uma frase tornam a tarefa de ensinar o computador a identificar padrões bastante complicada, sendo necessário a realização de tratamento dos dados que serão vistos nas próximas subseções.

#### 2.5.1.1 Tratamento dos dados

A etapa de tratamento dos dados realizada neste trabalho utiliza 5 técnicas que são: tokenização, remoção de *stopwords*, *stemming*, vetorização TF-IDF e *N*-gramas.

##### 2.5.1.1.1 Tokenização

O processo de tokenização é o processo de receber um texto e separar suas palavras de acordo com algum tipo de separação específica, gerando os *tokens*, que se referem às partes segregadas de determinado texto (PALMER, 2010). Este processo é a primeira etapa a ser realizada na extração de informação de textos, um dos separadores mais utilizados neste processo é o espaço em branco, onde este é descartado mantendo-se somente as palavras e outros delimitadores como sinais de pontuação, vírgulas, pontos e vírgulas, dentre outros.

Um exemplo da aplicação do processo de **tokenização** em uma frase ocorre da seguinte forma:

"Os queijos artesanais, são excelentes produtos! Uma verdadeira delícia."

Ao aplicar a *tokenização* na frase, tem-se o seguinte resultado:

```
["Os", "queijos", "artesanais", ",", "são", "excelentes", "produtos", "!",  
"Uma", "verdadeira", "delícia", "."]
```

Nesta simples frase, é possível verifica-se a presença de artigos, vírgulas e sinais de pontuação. De um modo geral, estes elementos em um pequeno texto são os que mais ocorrem podendo ser removidos pelo fato de serem considerados semanticamente irrelevantes para o entendimento do texto.

#### 2.5.1.1.2 Remoção de *stopwords*

Após a etapa de *tokenização*, em geral realiza-se a etapa de remoção das *stopwords* (ou palavras de parada - tradução livre do inglês) que como já dito são palavras consideradas “irrelevantes” em um determinado texto. Na língua portuguesa, as *stopwords* são compostas por artigos (a, o, as, os, um, uma, uns, umas), preposições, conjunções, dentre outros. Porém, ao analisar dados de redes sociais há a presença de outros elementos como URL's, caracteres especiais ([&| \* - +]), sinais de pontuação (.:!?), acentos (^~`), numerais ([0-9]) e símbolos que os acompanham como “R\$”, “\$”, “US\$”, “kg”, “km”. Todos são incluídos à esta listagem com intuito de remover o máximo de elementos que não trazem informação de maneira significativa. A normalização do texto para caixa baixa (*lowercase*) também é realizada, com intuito de diminuir a redundância de palavras que traduzem o mesmo significado, porém com grafias diferentes, fazendo com que o algoritmo as interprete como diferentes, como o caso dos termos: “queijo e Queijo”, que sintaticamente se referem à mesma coisa e se diferem apenas pela letra inicial onde uma é minúscula e a outra é maiúscula.

Na frase de exemplo usada na subseção anterior descrita abaixo:

```
["Os", "queijos", "artesanais", ",", "são", "excelentes", "produtos", "!",  
"Uma", "verdadeira", "delícia", "."]
```

há a presença de vários elementos considerados *stopwords* que são o caso de:

```
["Os", ",", "são", "!", "Uma", "."]
```

Nesta simples frase, cerca de 50% das palavras não traz nenhuma informação adicional, mostrando assim como esta etapa de remoção é importante para o processamento e, conseqüentemente, a qualidade dos dados no processo de mineração de informação. Aplicando todo o processo de normalização na frase de exemplo tem-se o seguinte resultado:

["queijos", "artesanais", "excelentes", "produtos", "verdadeira", "delicia"]

#### 2.5.1.1.3 Stemming

Outro tratamento bastante utilizado no processamento de linguagem natural é a técnica chamada *stemming*, que é responsável por transformar as flexões de uma palavra em um núcleo comum. Identificando o radical de uma palavra, faz-se com que palavras mais complexas morfológicamente sejam identificadas, decompondo-as em seu *stem* (tronco, unidade) invariante (HIPPISEY, 2010). Ao aplicar esta técnica ao resultado do processamento da remoção das *stopwords* presente na subseção dado por:

["queijos", "artesanais", "excelentes", "produtos", "verdadeira", "delicia"]

tem-se a redução dessas palavras aos respectivos radicais:

["queij", "artesan", "excel", "produt", "verd", "delic"]

dessa forma, há uma grande abstração no significado das palavras e uma redução clara no tamanho do vocabulário, pois variações de palavras como “queijeiro, queijaria, queijos”, “verdadeiro, verdadeiramente, verdade”, são todas reduzidas aos seus respectivos radicais, “queij” e “verdad”, impactando diretamente nos resultados das análises.

#### 2.5.1.1.4 Vetorização TF-IDF

Como já dito anteriormente, é muito comum ter palavras que não são tão relevantes em um determinado texto. Dessa maneira, a técnica TF-IDF (sigla do inglês *Term-Frequency Inverse-Document-Frequency*, que em tradução livre significa Frequência dos Termos - Inverso da Frequência do Documento) consiste em representar textos em formato de um vetor de termos para a determinação do peso que cada palavra possui em um documento (WILLIAM, 2019). Por exemplo, uma palavra que está presente em todas as frases de uma determinada base, não carrega consigo muita informação e por isso tem um peso baixo.

Baseado em William (2019), matematicamente o TF (Frequência do Termo) leva em consideração a frequência com que cada palavra **p** aparece em uma determinada frase **f**, sendo assim:

$$TF(p, f) = \text{QUANTIDADE DE VEZES QUE A PALAVRA } p \text{ APARECE NA FRASE } f \quad (2.1)$$

Já o IDF (Inverso da Frequência do Documento), se refere onde a palavra será ponderada quanto a sua capacidade de diferenciar as frases. Esse valor é o inverso da frequência de documento, ou seja, a frequência de documento que contém a palavra  $\mathbf{p}$ , logo  $FD(p)$  é dada por:

$$FD(\mathbf{p}) = \frac{\text{QUANTIDADE DE FRASES DA BASE QUE POSSUI A PALAVRA } \mathbf{p}}{\text{QUANTIDADE DE FRASES TOTAIS } N} \quad (2.2)$$

e por fim, o IDF é da forma:

$$IDF(\mathbf{p}) = \frac{N}{FD(\mathbf{p})} \quad (2.3)$$

Com o objetivo de evitar valores muito grandes de IDF, adiciona-se a função LOG ao resultado, evitando ainda a divisão por zero ao adicionar 1 ao termo do denominador. Sendo assim:

$$IDF(\mathbf{p}) = \text{LOG}\left(\frac{N}{FD(\mathbf{p}) + 1}\right) \quad (2.4)$$

Com isso, caso todos os documentos possuam uma determinada palavra  $\mathbf{p}$ , o valor do peso será  $\text{LOG}(N/N+1) \approx \text{LOG}(1) = 0$ . Dessa forma, esse parâmetro IDF permite que palavras que aparecem em muitos documentos tenham pesos menores, indicando assim, uma menor relevância se comparada a palavras com pesos maiores. Dessa forma, o valor final de cada posição dos vetores que representam as frases é dado por:

$$TFIDF(\mathbf{p}, \mathbf{f}) = TF(\mathbf{p}, \mathbf{f}) \cdot \text{LOG}\left(\frac{N}{FD(\mathbf{p}) + 1}\right) \quad (2.5)$$

#### 2.5.1.1.5 $N$ -gramas

Definido como uma sequência contígua de  $N$  elementos podendo estes serem caracteres, palavras, sílabas, fonemas e pares-base, os  $N$ -gramas são muito utilizados no processamento de linguagem natural por serem responsáveis pela obtenção de uma certa memória ao longo do processamento, realizando a verificação da composição de palavras em uma frase (JESÚS, 2019). Dessa maneira, se há interesse em palavras com  $N = 1$  dá-se o nome de unigrama, para  $N = 2$  bigrama ou digrama,  $N = 3$  um trigrama e, assim por diante.

Para a análise das características e hábitos de consumo do queijo artesanal no Brasil juntamente com a análise de conteúdo, a utilização desta técnica é de extrema importância, pois ao realizar a etapa de **tokenização**, cria-se automaticamente vários unigramas. Dessa

maneira, quando se procura pelo termo “doce de leite” no *tweet* fictício “Queijo com doce de leite é maravilhoso”, o algoritmo não consegue encontrar tal termo devido a dois fatores, sendo o primeiro causado pela presença de uma *stopword*, que neste caso é o “de”, que, devido ao processamento realizado no texto do *tweet*, após a remoção dos termos de parada, torna-se impossível encontrá-lo, sendo necessário pesquisar pelo bigrama “doce leite”. O outro fator se dá pelo processo de segmentação das palavras da frase com o delimitador espaço em branco realizado na etapa de *tokenização*, onde aplicando-o tem-se:

```
["queijo", "com", "doce", "de", "leite", "e", "maravilhoso", "."]
```

removendo as *stopwords*:

```
["queijo", "doce", "leite", "maravilhoso"]
```

é possível observar que com a remoção das *stopwords* e a *tokenização*, a relação “doce leite” é perdida, impossibilitando novamente encontrar “doce leite”.

Dessa forma, além de analisar unigramas, segmenta-se também os *tweets* em bigramas, de modo que, a frase de exemplo segmentada em bigramas já com a remoção de *stopwords* torna-se:

```
["queijo doce", "doce leite", "leite maravilhoso"]
```

e assim, é possível pesquisar por “doce leite” referindo-se ao derivado lácteo “doce de leite” e outros termos de interesse presentes no contexto dos queijos artesanais.

## 2.6 ANÁLISE DE SENTIMENTOS

Saber a opinião de outras pessoas ajuda e muito a maneira como as organizações podem abordar seus produtos. De acordo com Kelsey Group (2007), cerca de 81% dos usuários de internet já realizaram alguma pesquisa online sobre algum produto. Entre 73 - 87% dos usuários disseram que uma crítica/análise/revisão influenciaram na hora de comprar e que aproximadamente 99% dos consumidores estão dispostos a pagar por produtos avaliados com 5 estrelas ao invés de 4. Terra (2012) confirma os resultados desta pesquisa, em que fatores como a opinião de familiares, amigos, influenciadores digitais e sites de reclamação possuem forte influência na hora da compra.

Com o grande uso das redes sociais, entender o que os usuários dizem nestes ambientes é de extrema importância para entender o que esses consumidores buscam. Sendo assim, conhecida também por mineração de opinião, a análise de sentimentos, segundo Liu (2012), é uma área de estudo que busca entender e analisar as opiniões

expressas pelas pessoas sobre um determinado assunto, seja em textos, imagens, vídeos e etc. A análise de sentimentos visa identificar o sentimento apresentado a respeito de alguma entidade e descrever sentimentos e emoções descritas nos dados. Com o grande número de pessoas usando as mídias sociais, construir ferramentas que auxiliem organizações a entender melhor o perfil de consumo de seus produtos sobre determinados assuntos é crucial para atingir da melhor forma as expectativas dos consumidores.

Ao utilizar dados provenientes das redes sociais, um dos desafios mais comuns está relacionado à subjetividade presente em opiniões e em sentimentos. Dependendo da frase, a utilização de uma certa palavra pode fornecer um sentido completamente adverso causado pela utilização de uma linguagem informal com gírias, os famosos “memes”, palavras com duplo sentido, *emojis*, entre outros.

O Twitter (TWITTER, 2021) é uma das redes sociais mais populares e diversos estudos utilizam seus dados devido às características presentes nos mesmos. O primeiro ponto é referente à maior facilidade na obtenção de seus dados se comparado à outras redes sociais, outro fator é o de possuir conteúdos curtos que de uma forma geral expressa opiniões dos usuários contribuindo e muito para a utilização de técnicas de extração de informação como por exemplo mineração de dados, análise de sentimentos, processamento de linguagem natural, dentre outros. Diversos trabalhos utilizaram a rede social do Twitter como fonte de dados para a análise de sentimentos presente nos *tweets*. No trabalho de Andranik et al. (2010), a fim de prever o resultado da eleição federal alemã, os autores realizaram análises de conteúdo em mais de 100.000 publicações. Ao fim das eleições, puderam concluir que o resultado obtido pelas análises realizadas de fato refletiram no resultado final da eleição. Já em McGlohon et al. (2010), os autores analisaram comentários também do Twitter a fim de classificar produtos e comerciantes.

Uma das tarefas da análise de sentimentos é a de identificar a polaridade de um texto. A **polaridade** representa qual é o grau de positividade, neutralidade ou negatividade que um texto possui (BENEVENUTO et al., 2015). Neste contexto, a frase “Que dia lindo!” possui polaridade **positiva**, já a frase “Hoje é 22 de fevereiro de 2021” possui polaridade **neutra** e a frase “Que dia horrível” é de polaridade **negativa**. A tarefa de encontrar a polaridade de um texto é tratada como um problema de classificação utilizando uma abordagem de aprendizado supervisionado, onde é necessário desenvolver um conjunto de dados previamente rotulados para o treinamento do modelo de aprendizado de máquina. De forma geral, esta tarefa requer um grande esforço manual, pois na maioria dos problemas não se tem *datasets* específicos sendo necessário criá-los para a realização deste tipo de análise.

### 3 MATERIAL E MÉTODOS

#### 3.1 FLUXO DE TAREFAS

Para o desenvolvimento deste trabalho, três etapas principais foram necessárias: (i) coleta e armazenamento dos dados, (ii) processamento dos dados e (iii) pós-processamento dos dados e extração de informação. A Figura 8 exibe o fluxo de tarefas realizado pela ferramenta desenvolvida neste trabalho.

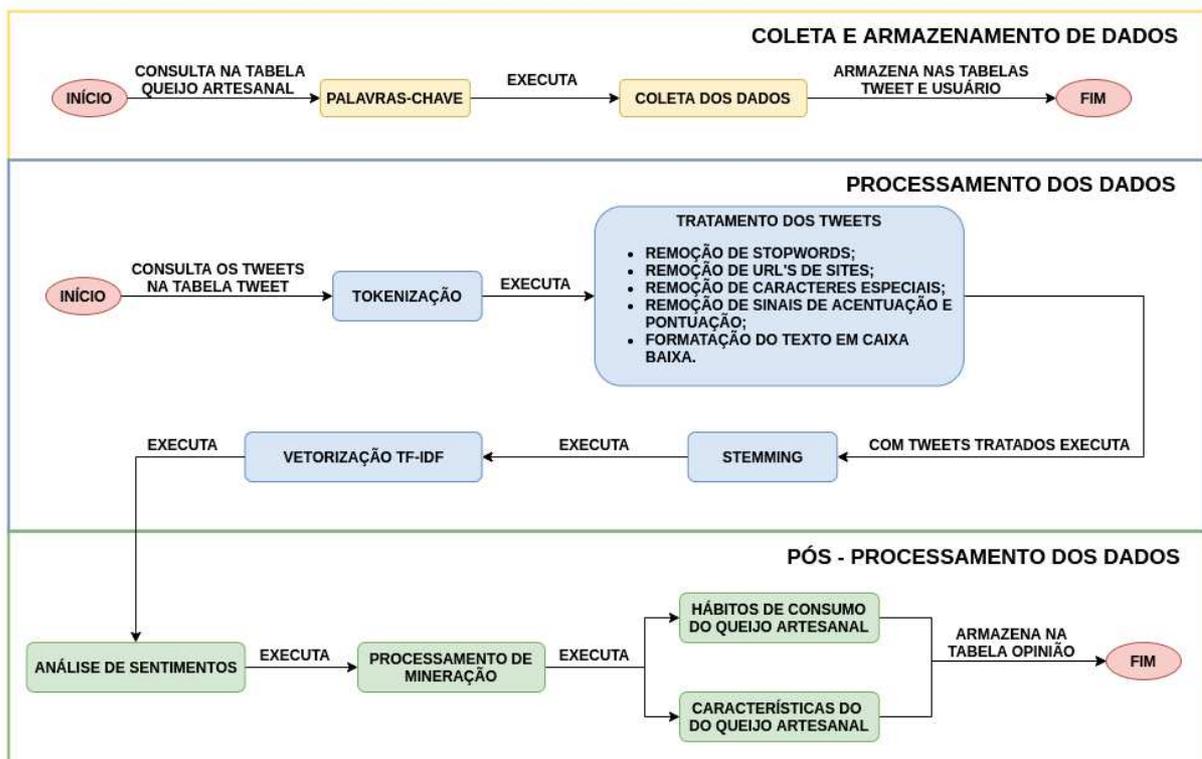


Figura 8 – Representação esquemática do fluxo de tarefas realizado neste trabalho. Em amarelo está representada a etapa de coleta e armazenamento dos dados que inicialmente realiza a consulta das palavras-chave no banco de dados e em seguida executa a coleta de *tweets* na plataforma do Twitter armazenando-os no banco de dados. Em azul é apresentada a etapa de processamento dos dados, que consulta os *tweets* armazenados no banco de dados e em seguida executa o processamento de *tokenização* nos *tweets*. Em seguida realiza o tratamento dos dados e o processo de *stemming*, finalizando com a vetorização utilizando TF-IDF. Em verde é exibido a etapa de pós-processamento, que utilizando os dados já vetorizados, realiza a análise de polaridade dos sentimentos armazenando os resultados no banco de dados. Por fim, é realizada a consulta dos *tweets* classificados como positivo para a identificação das características e hábitos de consumo do queijo artesanal armazenando ao final os resultados no banco de dados. **Fonte:** autor.

### 3.1.1 Etapa 1: Coleta e Armazenamento dos Dados

#### 3.1.1.1 Determinação do Conjunto de Busca

Para definir o espaço de busca para as análises deste trabalho, dividiu-se essa etapa em três fases. A primeira é referente à coleta e armazenamento dos dados do Twitter. Inicialmente, foi necessário desenvolver uma lista contendo diversas palavras-chave, que neste caso, foram os nomes dos queijos artesanais de interesse a serem consultados na plataforma do Twitter. Com a supervisão e orientação de especialistas da Embrapa, o desenvolvimento desta lista, se deu pela utilização de uma lista de nomes de queijos artesanais brasileiros que foram premiados no maior concurso nacional de queijos artesanais denominado “Prêmio Queijo Brasil”, que em 2018 contou com a participação de 479 queijos, sendo 179 os premiados e utilizados para a realização da nossa coleta podendo ser consultados no Apêndice A. Com tantas palavras-chave para a realização da coleta dos dados de queijos artesanais, foi necessário organizá-los em categorias como descrito nesta mesma listagem (COMERQUEIJO, 2019), com o intuito de agrupar os dados coletados, de modo mais eficiente, a fim de torná-los mais significativos. Essa estratégia foi adotada pelo fato de um determinado tipo de queijo possuir vários nomes diferentes, uma vez que são fabricados por produtores ou regiões diferentes, porém em sua essência se referem a um mesmo produto. Com isso, a Figura 9 exibe a quantidade de palavras-chave que cada categoria de tipo de queijo artesanal possui, que estão descritas na subseção 2.1.1.

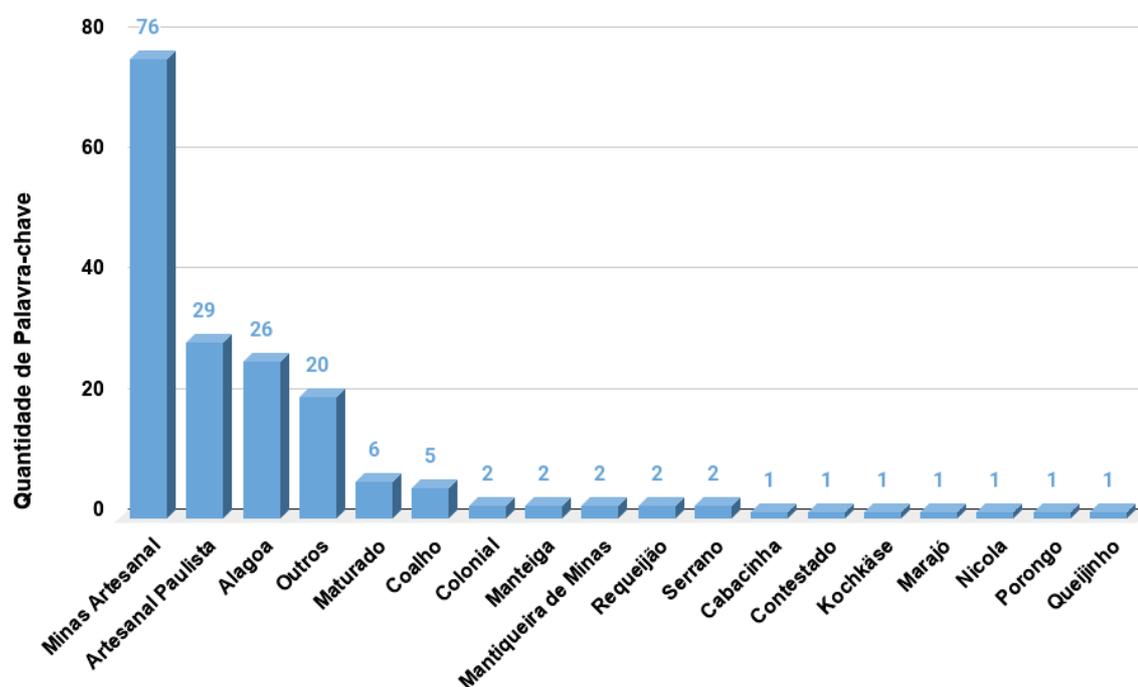


Figura 9 – Quantidade de palavras-chave por categoria de queijo. **Fonte:** autor.

### 3.1.1.2 Coleta de Dados do Twitter

De posse das palavras-chave, iniciou-se de fato a etapa de coleta dos dados da rede social. Para isso, foi desenvolvido um *script* na linguagem de programação  (R CORE TEAM, 2020), que importa a biblioteca *rtweet* (MW, 2019) responsável pela autenticação de desenvolvedor na plataforma do Twitter e em seguida realiza a coleta dos dados de acordo com a palavra-chave passada como parâmetro.

### 3.1.1.3 Estrutura dos dados do Twitter (*tweet*)

Ao realizar a coleta de dados do Twitter, é retornado um objeto denominado *tweet* com 92 campos textuais referentes ao usuário e ao conteúdo que o mesmo publicou. Nem todos os campos do objeto do *tweet* trazem informações relevantes neste trabalho e por esse motivo, foram selecionados 29 campos que podem ser visualizados no Apêndice B.

Devido às limitações presentes na API<sup>1</sup> gratuita do Twitter, que somente retorna dados de até 7 dias anteriores, foi necessário a configuração da execução automática do *script* de coleta dos dados semanalmente, com o principal objetivo de obter um conjunto de dados históricos para análises deste trabalho. Com tantas informações a serem armazenadas e o grande número de *tweets* coletados ao longo do tempo, foi necessário desenvolver um modelo de banco de dados capaz de armazenar de forma eficiente todos os *tweets* coletados, o que será melhor detalhado na próxima subseção.

### 3.1.1.4 Armazenamento dos Dados

Para armazenar os dados coletados do Twitter juntamente com todas as análises de interesse vindas do processamento dos mesmos da plataforma do “Observatório do Consumidor”, foi necessário realizar uma modelagem capaz de organizar da melhor forma todos os dados coletados e os que seriam gerados com o processamento das análises da ferramenta. Dessa forma, foi desenvolvido um banco de dados relacional utilizando o Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL (MYSQLLAB, 1995). Neste trabalho o MySQL foi utilizado pelo fato de ser um *software open source*<sup>2</sup>, possuir consistência, alta performance, confiabilidade e ser amplamente utilizado pelas indústrias (OFICINA, 2007).

<sup>1</sup> Do inglês *Application Programming Interface* e segundo Twitter (2021) permite amplo acesso aos dados públicos do Twitter que os próprios usuários escolheram compartilhar.

<sup>2</sup> Segundo Red Hat (2021) um *software open source* é um código que foi projetado para ser acessado abertamente pelo público, ou seja, qualquer pessoa pode visualizar, editar e distribuí-lo de acordo com sua necessidade.

### 3.1.1.4.1 Modelagem do Banco de Dados MySQL

Após o levantamento de todos os dados que seriam inseridos no SGBD do “Observatório do Consumidor”, foi desenvolvido o modelo de banco de dados relacional descrito na Figura 10.

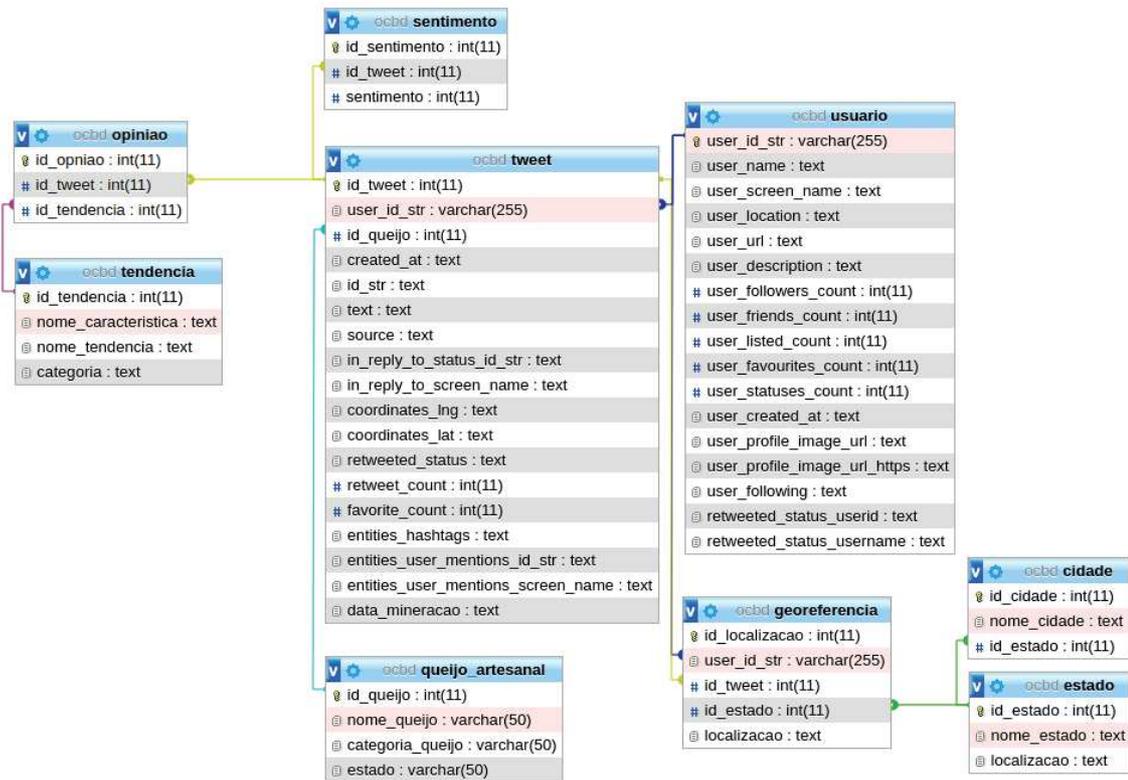


Figura 10 – Modelo do Banco de Dados Relacional da Plataforma do “Observatório do Consumidor”. **Fonte:** autor.

Ao todo foram desenvolvidas 9 tabelas. O fluxo de armazenamento dos dados começa da seguinte maneira: a plataforma do “Observatório do Consumidor” realiza uma consulta das palavras-chave armazenadas na tabela `queijo_artesanal`, que guarda informações como o nome do queijo, qual a sua categoria e a qual estado brasileiro o mesmo pertence, dessa forma, essa tabela é fixa e só é preenchida uma vez com os queijos de interesse. A tabela `tweet` é responsável por armazenar todas as informações referentes ao conteúdo publicado por um usuário que por sua vez é guardado na tabela `usuario`. Desta maneira, além dos campos já descritos presentes na tabela `tweet`, foi necessário criar mais um campo que é `data_mineracao` que guarda a data em que um `tweet` foi coletado a fim de otimizar consultas futuras pelos demais processamentos realizados pela ferramenta. A tabela `opiniaio` armazena os resultados obtidos após a etapa de identificação das características do queijo e hábitos de consumo que compõem a chamada tendência de consumo de interesse que está armazenada na tabela `tendencia`. A tabela `georeferencia` contém informações de localização geográfica do usuário e do `tweet` quando

estas informações são disponibilizadas pelos usuários. Por fim, as tabelas **estado** e **cidade** foram criadas contendo todos estados e cidades do Brasil para facilitar o processamento da localização do usuário a ser exibida na ferramenta.

### 3.1.2 Etapa 2: Processamento dos Dados

#### 3.1.2.1 Mineração de Dados Textuais

Após todo o armazenamento referente à etapa de coleta dos dados do Twitter, inicia-se de fato a etapa de mineração em busca de informações que estão presentes nos dados coletados semana a semana. Inicialmente, realiza-se a consulta no banco de dados de todos os *tweets* que foram armazenados referentes à semana em questão. Dessa maneira, os *tweets* passam por todo o processamento de linguagem natural descrito anteriormente, realizando-se a técnica de **tokenização** de todos os *tweets*. Em seguida, remove-se todas as **stopwords** de interesse e normaliza-se todos os *tweets* em *lowercase* (letras minúsculas). Feito isso, com os *tweets* tratados, inicia-se a etapa de **stemming** reduzindo todas palavras de todos os *tweets* ao seu respectivo radical. Em seguida, realiza-se o último passo antes da análise de sentimentos que é a vetorização dos *tweets* utilizando o TF-IDF, verificando as palavras mais relevantes em cada *tweet*. Com todos os *tweets* devidamente tratados, inicia-se de fato a análise de sentimentos atribuindo aos *tweets* a polaridade do sentimento presente nos comentários.

### 3.1.3 Etapa 3: Pós-Processamento dos Dados

#### 3.1.3.1 Interpretação dos Dados

Finalmente, na etapa 3 foram implementados algoritmos para a análise de sentimentos dos *tweets*, utilizando um aprendizado de máquina supervisionado para realização da classificação dos sentimentos dos *tweets* em negativos, neutros e positivos. Com os *tweets* classificados, iniciou-se o processamento de clusterização que tem a função de preparar os dados para a identificação das tendências de consumo, respondendo quais são as características e hábitos de consumo do queijo artesanal no Brasil.

## 3.2 MODELO DE ANÁLISE DE SENTIMENTOS

Para o desenvolvimento do modelo de análise de sentimentos e processamento de linguagem natural, a linguagem de programação escolhida foi Python (FOUNDATION, 2021) por ser amplamente utilizada para esta finalidade em conjunto com as bibliotecas Scikit-Learn (PEDREGOSA et al., 2011a), NLTK (LOPER; BIRD, 2002), Pandas (MCKINNEY, 2010), dentre outras.

### 3.2.1 Ensemble Voting Classifier

Os métodos de *ensemble* (em tradução livre do inglês, conjunto), referem-se à combinação de classificadores de aprendizado de máquina conceitualmente diferentes a fim de utilizar um voto majoritário ou as probabilidades médias previstas para previsão dos rótulos das classes (PEDREGOSA et al., 2011b). Dessa forma, este classificador se mostra bastante útil para equilibrar fraquezas individuais de vários classificadores, aumentando o desempenho da atividade. Neste trabalho os classificadores utilizados no *Voting Classifier* foram: `Logistic Regression`, `Multinomial Naive Bayes`, `OneVsRestClassifier` e `OneVsOneClassifier`.

#### 3.2.1.1 Classificadores

##### 3.2.1.1.1 Regressão Logística

Ao contrário do que o nome diz, a regressão logística é um modelo linear e é utilizado para classificação ao invés da regressão. Neste modelo, utiliza-se técnicas estatísticas com o objetivo de fornecer a partir de um determinado conjunto de observações, predições de valores de uma variável categórica podendo esta ser binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias (GONZALEZ, 2018). Parecido com a regressão linear, a regressão logística necessita da aplicação de pesos que se ajustam aos dados de treinamento do algoritmo a fim de obter a melhor curva ao invés da melhor reta como proposto na regressão linear. A regressão logística calcula uma razão de probabilidade da variável alvo, que posteriormente é convertida em uma variável de base logarítmica, permitindo assim a classificação com base na aproximação de um dos valores (WITTEN; FRANK, 2005).

A equação utilizada para a realização da classificação do tipo de polaridade utilizando o classificador de Regressão Logística é dada da seguinte forma (SOUZA, 2017):

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

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

##### 3.2.1.1.2 Naïve Bayes

Os métodos Naïve Bayes pertencem a um conjunto de algoritmos de aprendizagem supervisionada que se baseiam na aplicação do teorema de Bayes com a suposição “ingênua” (*naïve*) de independência condicional entre cada par de características dado o valor da

variável de classe. O teorema de Bayes (PEDREGOSA et al., 2011b), afirma a seguinte relação, dada variável de classe  $y$  e o vetor de recurso dependente  $x_1$  através de  $x_n$ :

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)} \quad (3.2)$$

Utilizando a condição de independência tem-se que:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y) \quad (3.3)$$

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

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

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

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

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i|y) \quad (3.6)$$

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

A variação do Naïve Bayes utilizado neste trabalho é denominada Multinomial Naïve Bayes, e implementa tal algoritmo para dados distribuídos de forma multinomial sendo uma das variações clássicas utilizadas para classificação de textos. A distribuição é parametrizada por vetores  $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$  para cada classe  $y$ , sendo  $n$  o número de características e  $\theta_{yi}$  a probabilidade  $P(x_i|y)$  da característica  $i$  aparecer em uma determinada amostra pertencente a classe  $y$ .

Os parâmetros  $\theta_{yi}$  são estimados através de uma versão suavizada de probabilidade máxima, ou seja, contagem de frequência relativa, dada por:

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

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

### 3.2.1.1.3 OneVsRest e OneVsOne

O classificador chamado de um contra todos (*OneVsRest*), segundo Pedregosa et al. (2011b), se baseia em ajustar um classificador por classe. Dessa forma, para

cada classificador, ajusta-se a classe atual em relação a todas as demais classes. Este classificador é bastante eficiente e outra vantagem é sua interpretabilidade, como cada classe é representada por somente um classificador, torna-se possível obter conhecimento sobre determinada classe inspecionando seu respectivo classificador.

A estratégia chamada de um contra um (*OneVsOne*), consiste em ajustar um classificador por par de classes (PEDREGOSA et al., 2011b). No momento da previsão, a classe que recebeu mais votos é selecionada. Uma vez que necessita do ajuste de  $\frac{n\_classes \cdot (n\_classes - 1)}{2}$  classificadores, este método é geralmente mais lento do que *OneVsRest*, devido à sua complexidade  $\mathbf{O}(n\_classes^2)$ . No entanto, este método pode ser vantajoso para algoritmos como algoritmos de *kernel* que não escalam bem com `n_amostras`. Isso ocorre porque cada problema de aprendizagem individual envolve apenas um pequeno subconjunto de dados, enquanto, com um contra o resto, o conjunto de dados completo é usado  $n\_classes$  vezes.

### 3.2.1.2 Votação por Decisão da Maioria

Neste tipo de votação, a estratégia utilizada se baseia em rotular à classe que será prevista, o rótulo referente ao que a maioria dos classificadores previram. Dessa forma, suponha que haja 3 classificadores com suas respectivas classificações para uma determinada amostra iguais a:

classificador 1  $\rightarrow$  *classe1*  
 classificador 2  $\rightarrow$  *classe1*  
 classificador 3  $\rightarrow$  *classe2*

o classificador com o parâmetro `voting = hard`, classificaria tal amostra como sendo da classe 1, pois a maioria dos classificadores rotularam este resultado. Em caso de empate, o classificador de votação majoritária selecionaria a classe com base na ordem de classificação crescente. Por exemplo, no seguinte cenário:

classificador 1  $\rightarrow$  *classe2*  
 classificador 2  $\rightarrow$  *classe1*

o rótulo de classe 1 seria atribuído à determinada amostra.

## 3.2.2 Base de Treinamento do Modelo

Para a construção deste *dataset* de treinamento, utilizamos 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. Para a pré-classificação do conjunto de 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.

Este processo foi feito para se ter uma base de treinamento bem acurada, dado que não foi encontrado nenhum *dataset* específico para análise de sentimentos de *tweets* no idioma português. 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. A Figura 11 mostra em detalhes este desbalanceamento presente nos dados de treinamento.

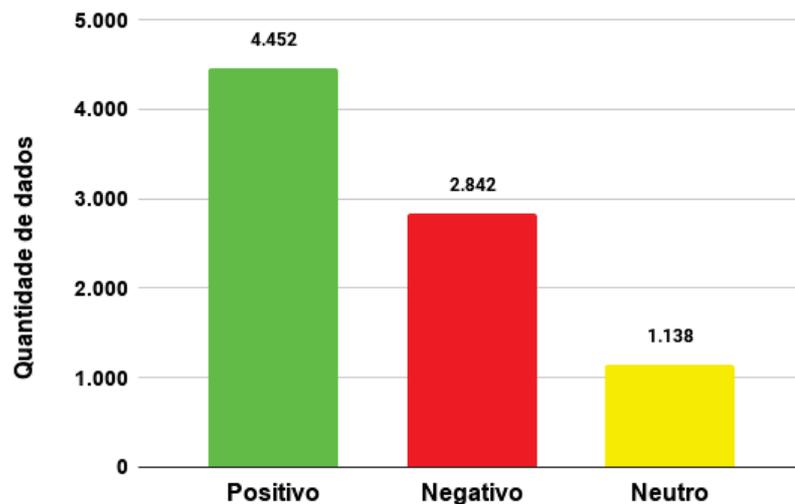


Figura 11 – Desbalanceamento presente nos dados da base de treinamento do modelo. **Fonte:** autor.

### 3.2.3 Validação Cruzada

Para a validação do modelo foi utilizada uma técnica de avaliação estatística chamada de validação cruzada, que tem como principal objetivo avaliar a capacidade de um modelo generalizar um conjunto de dados verificando o quão preciso é o modelo ao receber um novo conjunto de dados. A técnica de validação cruzada usada neste trabalho foi a *K-fold*, em que a ideia principal consta em dividir o conjunto de dados de treinamento em  $K$  partes, sendo uma parte para treino e outra para teste. Desta forma, o modelo segue um processo iterativo de aprendizado. Para ilustrar esse processo, é escolhido um conjunto contendo  $N$  amostras. Em seguida esse conjunto é dividido em  $K$  subconjuntos em que  $K \geq 1$ . Desses  $K$  subconjuntos,  $K - 1$  são escolhidos de forma randômica para treino e o restante para teste. Esse processo é realizado  $K$  vezes sempre com um conjunto diferente de treino e teste. Com este processo a validação *K-fold* consegue aumentar o desempenho do modelo reduzindo a sensibilidade ao particionar os dados. A Figura 12 esquematiza um exemplo de conjunto de dados dividido em  $K = 5$  subconjuntos a cada iteração do processo de validação cruzada.

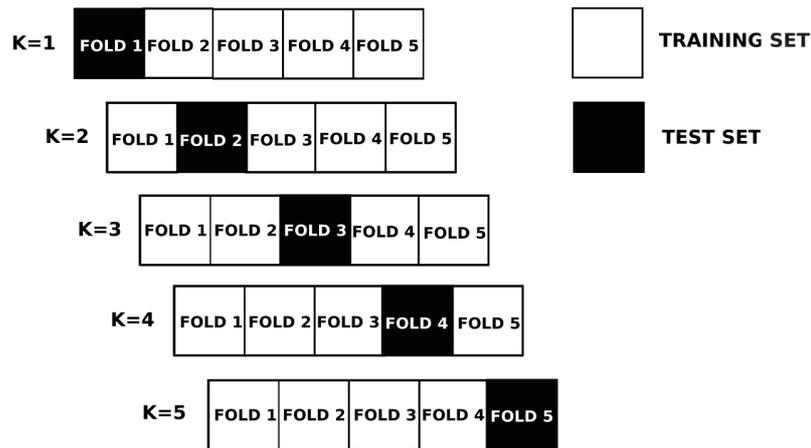


Figura 12 – Representação esquemática do diagrama de validação cruzada com  $k = 5$ .  
**Fonte:** Baseado em (SOARES et al., 2017).

### 3.2.4 Métricas de Avaliação

Neste trabalho, foram implementadas quatro métricas de validação com o intuito de verificar quão bem o algoritmo classifica um conjunto de dados já classificado, verificando acertos e erros. O coeficiente de determinação  $R^2$ , é uma medida descritiva da qualidade do ajuste obtido, mensurando quão próximos os dados estão da linha de regressão ajustada. Ele fornece uma indicação da qualidade do ajuste e, portanto, uma medida de quão bem as amostras não vistas podem ser previstas pelo modelo. Seus valores variam de 0 a 1 e quanto mais próximo de 1 mais explicativo é o modelo.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.8)$$

em que  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  e  $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$ .

O MSE (do inglês *Mean Squared Error*), dado pela Equação 3.9, é responsável por calcular o erro médio quadrático, uma métrica de risco correspondente ao valor esperado do erro ou perda quadrática (quadrática). Quanto mais próximo de 0 melhor.

$$MSE(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 \quad (3.9)$$

em que  $\hat{y}_i$  é o valor predito da  $i$ -ésima amostra e  $y_i$  é o valor real no total de amostras.

A raiz do erro médio quadrático (RMSE, do inglês *Root Mean Squared Error*) é a raiz quadrada do MSE, que aplicada na Equação 3.9 tem-se a Equação 3.10, portanto valores próximos de 0 também são esperados.

$$RMSE(y_i, \hat{y}_i) = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2} \quad (3.10)$$

O erro absoluto mediano (MAE, do inglês *Mean Absolute Error*), apresentado na equação 3.11, é particularmente interessante pois é robusto para tratamento de *outliers*. A perda é calculada tomando a média de todas as diferenças absolutas entre o dado correto e a previsão.

$$MAE(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i| \quad (3.11)$$

A matriz de confusão (exemplificada na Figura 13), é uma forma de avaliar o desempenho de uma classificação utilizando critérios quantitativos. Dessa forma, segundo Rosenfield e Fitzpatrick (1986), a matriz de confusão é uma matriz quadrada de números, que expressa a relação do número de unidades da amostra de uma determinada categoria, e a categoria verdadeira que esses dados pertencem durante o processo de classificação. Em geral, as colunas representam os dados de referência e as linhas, a classificação resultante é obtida através de algum método de classificação.

		ACTUAL VALUES	
		POSITIVE (1)	NEGATIVE (0)
PREDICTED VALUES	POSITIVE (1)	TP	FP
	NEGATIVE (0)	FN	TN

Figura 13 – Matriz de Confusão. **Fonte:** Baseado em Narkhede (2018).

Dessa forma, os resultados podem ser resumidos em quatro valores que são:

- Verdadeiro positivo (true positive - TP): significa que a classe prevista e a observada originalmente foram classificadas corretamente como positivos;
- Falso positivo (false positive - FP): significa que a classe predita retornou positivo, mas a classe correta era negativa;
- Negativo verdadeiro (true negative - TN): os valores preditos e observados foram classificados corretamente como negativos;
- Falso negativo (false negative - FN): representa que o valor predito resultou na classe negativa, mas o correto seria classificar como positivo.

### 3.3 ANÁLISE DE CONTEÚDO

#### 3.3.1 Análise das Características e Hábitos de Consumo do Queijo Artesanal

Para identificar as tendências de consumo do queijo artesanal no Brasil, neste trabalho foram desenvolvidos dois conjuntos de palavras-chave, sendo um para identificar quais são as características do queijo artesanal e outro para os seus hábitos de consumo. Com base na opinião dos pesquisadores da Embrapa, foram criadas 15 categorias, sendo 12 para características do queijo artesanal e 3 para hábitos de consumo.

Assim sendo, uma categoria bastante importante de ser analisada era a do **Preço**, e ela foi responsável por identificar nos *tweets* o que era comentado sobre o preço dos queijos artesanais. A categoria **Tamanho** foi criada com intuito de avaliar os aspectos físicos com relação ao tamanho dos queijos nas publicações. Para tentar encontrar informações referentes ao sabor dos queijos nas publicações, a categoria **Sabor** foi criada e características como suave, agradável e gosto forte foram alguns termos de interesse. A categoria **Produção**, foi desenvolvida para identificar nas publicações dos usuários informações referentes às origens e modos de fabricação do queijo ou até mesmo em como é utilizado em determinados pratos ou receitas. A categoria **Maturado** possui atributos que fazem referência ao processo de maturação dos queijos, identificando se o queijo mencionado é maturado ou não, se é mais ou menos, dentre outros aspectos. A categoria **Textura** foi criada para identificar o que os usuários comentavam com relação à textura do queijo, verificando se o mesmo era cremoso, seco, mole, etc.

A categoria **Selo Arte**, foi desenvolvida para identificar se nas publicações sobre o queijo artesanal, os usuários diziam algo sobre o selo criado pelo Governo para identificar produtos seguros e de qualidade. A categoria **Composição** foi criada com o objetivo de identificar alguns produtos que são adicionados ao queijo artesanal em seu processo de fabricação. A categoria **Fungos** foi desenvolvida para identificar publicações que falavam sobre a presença de fungos ou mofos naturais que são bastante presentes em determinados queijos. Para analisar o quesito cor, foi criada a categoria **Cor** que identifica as cores que um determinado queijo mencionado em uma publicação possui. A categoria **Odor** extrai as informações referentes ao odor, verificando atributos como cheiro bom ou ruim, o aroma, entre outros atributos que definem as características dos queijos. Para identificar informações sobre a embalagem, a categoria **Embalagem** possui atributos que tentam identificar o modo como os queijos são embalados para o comércio.

Os termos que descrevem a categoria de **Acompanhamentos** foram responsáveis por identificar produtos que além de serem mencionados nas publicações são muito consumidos em conjunto com os queijos artesanais. A categoria **Receitas** foi desenvolvida com intuito de identificar pratos e receitas culinárias que possuem o queijo como um dos ingredientes. A categoria **Bebidas**, finaliza o grupo que define os hábitos de consumo e foi desenvolvida

para identificar as bebidas de interesse e que foram mencionadas juntamente com o consumo dos queijos artesanais. As bebidas por si só são acompanhamentos dos queijos, porém foi separada em uma categoria específica com o objetivo de deixar os resultados mais segmentados e mais simples de serem analisados.

A Tabela 1 mostra os atributos referentes às características e hábitos de consumo do queijo artesanal que foram utilizadas para as análises, sendo construída com base na opinião de especialistas da Embrapa Gado de Leite, representantes da Empresa de Assistência Técnica e Extensão Territorial (EMATER) e pela análise de palavras mais frequentes nas chamadas nuvens de palavras (do inglês, *wordclouds*). Um exemplo de *wordcloud* no contexto desta pesquisa pode ser visto na Figura 14.

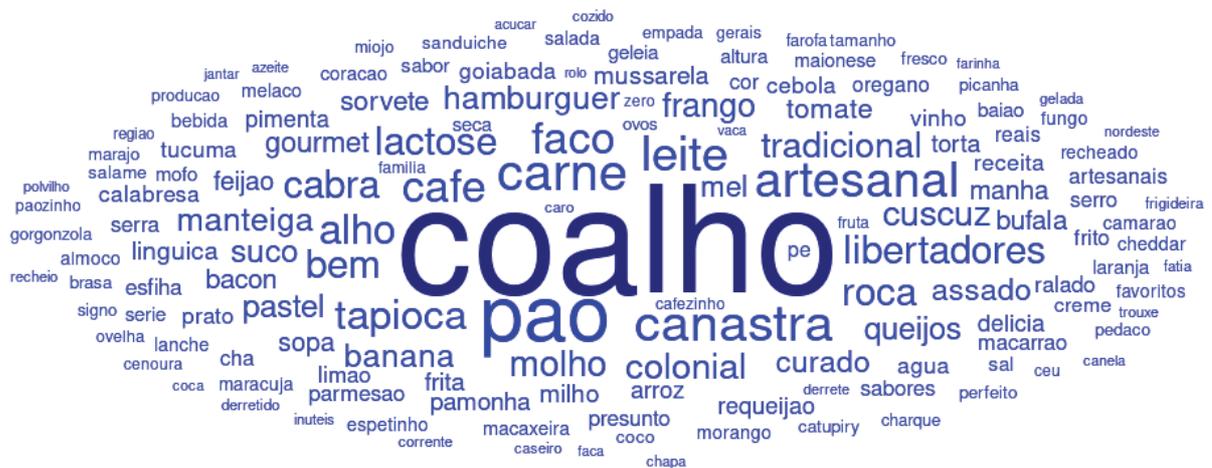


Figura 14 – Nuvem das palavras mais citadas ao longo das semanas de estudo (30/04/2020 a 18/02/2020). **Fonte:** autor.

Tabela 1 – Características e hábitos de consumo dos queijos artesanais analisados nos *tweets*. Os atributos foram construídos e minerados sem os sinais de acentuação e cedilhas devido ao processamento de linguagem natural realizar a remoção dos mesmos.

	<b>Categorias</b>	<b>Atributos</b>
<b>Características do queijo artesanal</b>	Preço	queijo barato, queijo caro, preco baixo, preco alto, justo.
	Tamanho	queijo pequeno, queijo medio, queijo grande, queijo padrao.
	Sabor	suave, sabor suave, agradavel, desagradavel, enjoativo, salgado, amargo, acido, picante, azedo, ardido, gosto forte, frutado, sal, saudavel, frescor, terroso, adocicado.
	Produção	artesanal, familiar, tradicional, natural.
	Maturado	maturado, mais maturado, menos maturado, pouco maturado, muito maturado, curado, fresco.
	Textura	cremoso, macio, massa macia, firme, duro, fragil, pastoso, pegajoso, granuloso, arenoso, fibroso, cristalino, seco, umido, oleoso, gorduroso, olhaduras, mole, fechada, elastica, amanteigada, untuoso.
	Selo Arte	higiene, qualidade, certificacao, inspecao, identidade, etiqueta caseina, selo arte.
	Composição	puro, temperado, oregano, ervas finas, defumado, lombinho, tomate seco, beterraba, urucum, tomilho, alecrim, pimenta rosa, condimento, especiarias, cheiro verde, manjericao, salsinha, cebola.
	Fungos	fungo branco, fungo preto, fungo cinza, contaminacao, mofo azul, mofo, mofo natural.
	Cor	branco, amarelo, amarelado, amarelo claro, amarelo escuro, amarelo esverdeado, manchado, pintado, amarelo palha, queijo azul, alaranjado.
	Odor	ausente, pronunciado, caracteristico, vegetal, rancoso, cheiro forte, aroma, cheiro ruim, cheiro bom, queijo fedorento.
	Embalagem	vacuo, cestas, embalagem, caixinha, palha, plastico, papel.
<b>Hábitos de consumo do queijo artesanal</b>	Bebidas	cafe, cha, leite, chimarrao, suco, limonada, refrigerante, guarana, coca cola, cachaca, vinho, tinto, cerveja, espumante, champagne, uisque, gim, tequila, vodka, caipirinha, conhaque, pinga, corote.
	Acompanhamentos	doce, doce leite, azeite, geleia, pao, biscoito, torrada, marmelada, azeitona, mel, melaco, melado, banana, pamonha, goiabada, uva, salaminho, salame, tabua, frios frito, assado, ralado, damasco seco, nozes, avela, carne seca, pimenta, batata, petiscos, castanha, calabresa, linguica, ovo codorna, limao, torresmo, bolacha, bolo, ameixa, morangos, peras, tucuma, bacon, mandioca, charque, carne, frango.
	Receitas	baiao dois, tapioca, feijao tropeiro, cartola, macarronada, macarrao, arroz, fondue, sanduiche, tira gosto, omelete, empanado, pao queijo, massas, lasanha, pizza, carne sol, escondidinho, hamburguer, misto quente, escrachete, bife parmegiana, pastel queijo, panqueca, crepe, risoto, creme queijo, pudim queijo, torta queijo, mousse queijo, almondegas, romeu julieta, farofa.

**Fonte:** autor.

Para a realização da extração das informações mencionadas acima dos *tweets* coletados, realizou-se inicialmente a consulta de todos os *tweets* classificados com a polaridade positiva, com o objetivo de realizar a análise de conteúdo e retornar dessa forma um resultado contendo opiniões positivas sobre as características e hábitos de consumo do queijo artesanal no Brasil.

### 3.3.2 Análise de Consumidores Reais e Potenciais Consumidores dos Queijos Artesanais

Com intuito de analisar os usuários do Twitter que disseram consumir ou estarem com vontade de consumir os queijos artesanais ou pratos culinários que possuem o queijo como ingrediente, foram criadas duas categorias de atributos, uma denominada **Consumidores reais** com a presença de termos que indicam que o usuário/consumidor consumiu um determinado produto e outra denominada **Potenciais Consumidores**, que identifica nos *tweets* a vontade dos usuários/consumidores consumirem um determinado produto. Ambas categorias foram criadas com base em análises realizadas nos comentários dos usuários. A Tabela 2 mostra em detalhes os atributos que foram minerados nos *tweets*.

Tabela 2 – Características do consumo analisados nos *tweets*. Os atributos foram construídos e minerados sem os sinais de acentuação e cedilhas devido ao processamento de linguagem natural realizar a remoção dos mesmos.

<b>Consumidores reais</b>	como, comer, comi, comendo, beber, bebi, bebendo, consumo, consumir, consumindo, ingerir, comprar, comprei, comprado, tomar, tomei, tomo, tomando, beliscar, engolir, devorar, devorei, devorando, saborear, provar, petiscar, papar, papei, almocar, almocei, jantar, jantei, almocando, jantando, rangar, rangei, fazer, ficou, fiz, acabei, pedi.
<b>Potenciais consumidores</b>	vontade, quero, queria, pudesse, poderia, comeria, comia, pensando, querendo.

**Fonte:** autor.

## 4 RESULTADOS E DISCUSSÃO

### 4.1 PANORAMA GERAL

Nesta seção será mostrado uma visão geral sobre a quantidade de *tweets* ao longo do tempo, as categorias dos tipos de queijos artesanais que mais foram mencionados nas publicações e os estados brasileiros que mais publicaram.

#### 4.1.1 Análise Temporal

Ao todo foram coletados 82.959 *tweets* entre 30 de abril de 2020 e 18 de fevereiro de 2021, totalizando quarenta e duas semanas. A Figura 15 mostra o panorama da quantidade de *tweets* coletados ao longo das semanas<sup>1</sup> do período analisado.

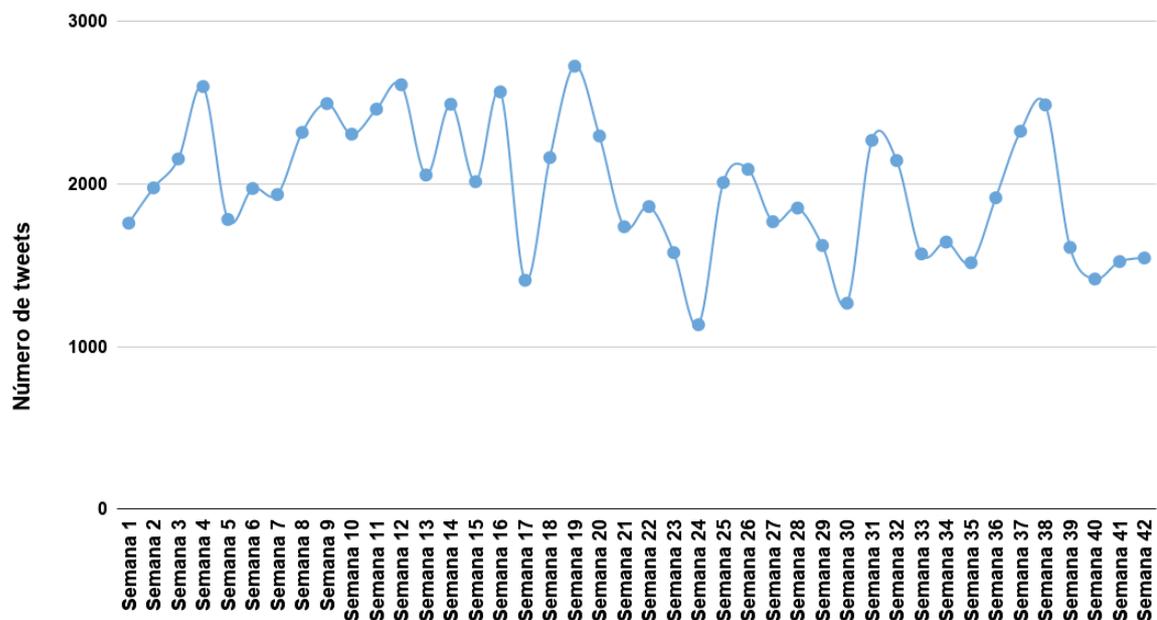


Figura 15 – Quantidade de *tweets* ao longo das semanas. **Fonte:** autor.

<sup>1</sup> Semana 1 - 30/04 a 7/05, Semana 2 - 7/05 a 14/05, Semana 3 - 14/05 a 21/05, Semana 4 - 21/05 a 28/05, Semana 5 - 28/05 a 4/06, Semana 6 - 4/06 a 11/06, Semana 7 - 11/06 a 18/06, Semana 8 - 18/06 a 25/06, Semana 9 - 25/06 a 2/07, Semana 10 - 2/07 a 9/07, Semana 11 - 9/07 a 16/07, Semana 12 - 16/07 a 23/07, Semana 13 - 23/07 a 30/07, Semana 14 - 30/07 a 6/08, Semana 15 - 6/08 a 13/08, Semana 16 - 13/08 a 20/08, Semana 17 - 20/08 a 27/08, Semana 18 - 27/08 a 3/09, Semana 19 - 3/09 a 10/09, Semana 20 - 10/09 a 17/09, Semana 21 - 17/09 a 24/09, Semana 22 - 24/09 a 1/10, Semana 23 - 1/10 a 8/10, Semana 24 - 8/10 a 15/10, Semana 25 - 15/10 a 22/10, Semana 26 - 22/10 a 29/10, Semana 27 - 29/10 a 5/11, Semana 28 - 5/11 a 12/11, Semana 29 - 12/11 a 19/11, Semana 30 - 19/11 a 26/11, Semana 31 - 26/11 a 3/12, Semana 32 - 3/12 a 10/12, Semana 33 - 10/12 a 17/11, Semana 34 - 17/12 a 24/12, Semana 35 - 24/12 a 31/12, Semana 36 - 31/12 a 7/01, Semana 37 - 7/01 a 14/01, Semana 38 - 14/01 a 21/01, Semana 39 - 21/01 a 28/01, Semana 40 - 28/01 a 4/02, Semana 41 - 4/02 a 11/02, Semana 42 - 11/02 a 18/02.

Das 179 palavras-chave desenvolvidas para a coleta de dados, apenas 80 retornaram *tweets* ao longo das coletas realizadas. Tal fato pode ser explicado pela existência de uma enorme variedade de queijos artesanais que por serem tão específicos, a população brasileira que utiliza o Twitter não os conhece o bastante para mencioná-los em suas publicações. Sabendo disso, é possível verificar que no geral, o quantitativo de *tweets* sofre diversas oscilações semana após semana, porém, mantém um patamar médio de aproximadamente 1.997 *tweets* ao longo das semanas. A partir da semana 20, é possível verificar que a quantidade de *tweets* publicados fica abaixo do valor médio e tal redução repentina do número de *tweets* pode estar relacionado com a volta parcial da rotina dos brasileiros após um longo período de isolamento social, devido à pandemia da COVID-19.

#### 4.1.2 Categorias de Tipos de Queijos Artesanais mais Citados no Twitter

A Figura 16, mostra as categorias de queijos mais comentados no Twitter no Brasil.

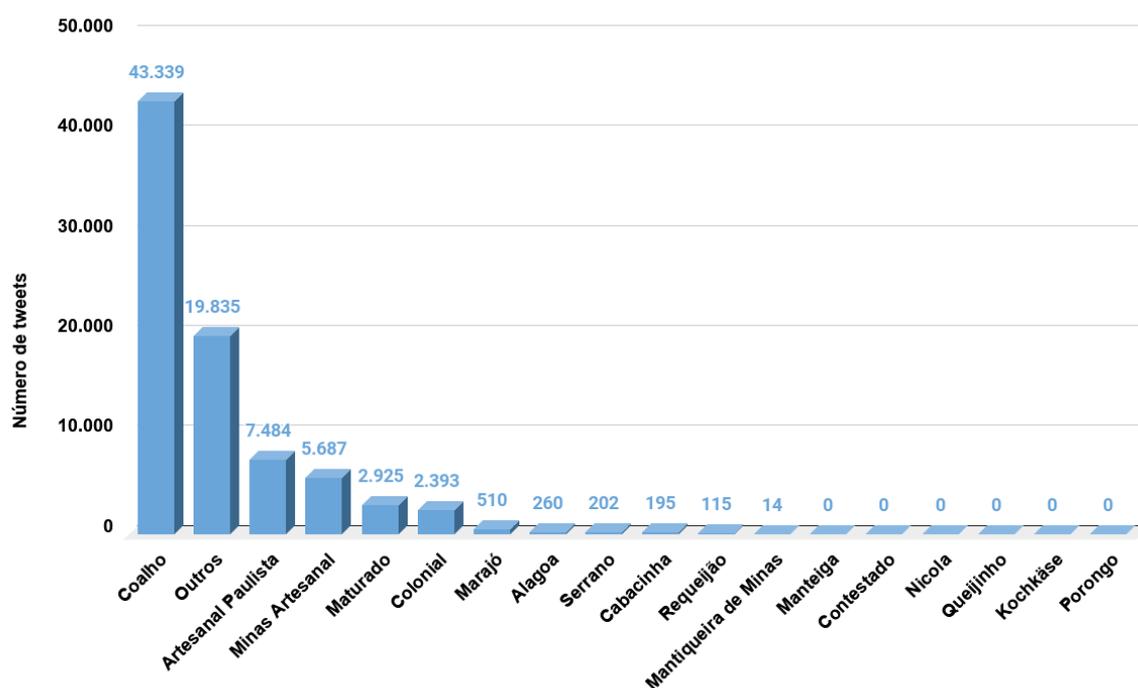


Figura 16 – Categorias de queijos artesanais mais citados no Brasil no período da pesquisa.  
**Fonte:** autor.

A categoria que representa o queijo Coalho, foi a que mais se destacou no período em análise com aproximadamente 52,2% do total dos *tweets* coletados. No entanto, a palavra coalho possui certa limitação pelo fato de abranger diversos possíveis contextos. Logo, em certas ocasiões ela pode fazer referência ao coalho utilizado para coagular o leite no processo de produção do queijo ou estar relacionada ao próprio queijo Coalho, podendo este ser artesanal ou industrial. A categoria Outros representa os produtos que

não se encaixaram em nenhuma das categorias disponíveis, Artesanal Paulista, Minas Artesanal, Maturado e Colonial.

#### 4.1.3 Estados Brasileiros que mais Comentaram Sobre Queijos Artesanais

Para identificar quais regiões brasileiras mais publicaram ao longo do tempo de estudo em questão, foi realizado um processamento no campo `location` do *tweet*, com o intuito de identificar e contabilizar as localizações válidas dos estados de cada usuário. Dessa forma, a Figura 17 exhibe os resultados obtidos.

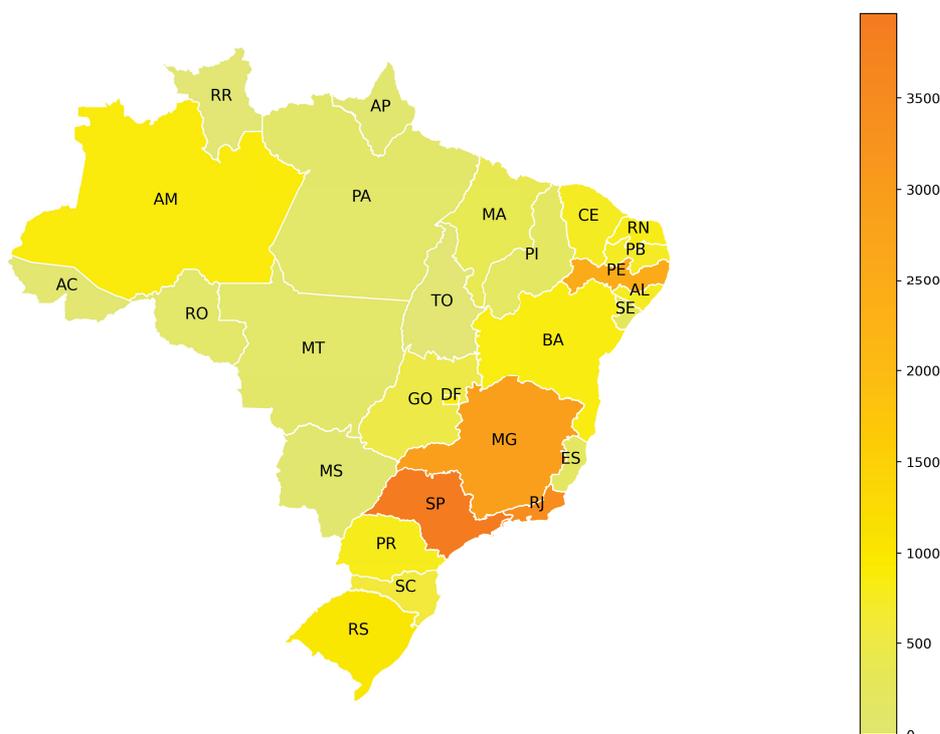


Figura 17 – Publicações por estados brasileiros sobre queijo artesanal ao longo das semanas de estudo (30/04/2020 a 18/02/2020). **Fonte:** autor.

Ao todo, cerca de 22.635 *tweets* continham a informação de localizações válidas, representando cerca de 27,3% do total de *tweets* analisados. Dessa maneira, é possível verificar que os 10 estados que mais publicaram foram os de São Paulo, Rio de Janeiro, Minas Gerais, Pernambuco, Rio Grande do Sul, Amazonas, Distrito Federal, Bahia, Rio Grande do Norte e Paraná, que juntos representaram cerca de 79,4% do total de localizações válidas.

Nos estados de São Paulo e Rio de Janeiro os tipos de queijos artesanais mais citados foram o *Coalho* e *Outros* que juntos representaram respectivamente 75,2% e 60,5% das publicações. No estado de Minas Gerais as categorias *Outros* e *Minas Artesanal* foram os de maior destaque totalizando 62,1% de publicações. Em Pernambuco, o queijo *Coalho* foi o mais mencionado com 91,6% das publicações, sendo o mais representativo

deste estado. No Rio Grande do Sul, os queijos mais citados foram o Colonial e Coalho que, juntos representaram 68,1% das publicações. No estado do Amazonas e no Distrito Federal o queijo Coalho representou respectivamente 86,3% e 51,4% das publicações. Na Bahia, no Rio Grande do Norte e Paraná novamente o queijo mais citado foi o Coalho com 79,2%, 72,6% e 41,3% das publicações sobre os queijos artesanais, respectivamente. É interessante notar que o queijo Coalho obteve uma representatividade bastante expressiva em diversos estados brasileiros, mesmo em regiões em que não é descrito como queijo típico regional, entretanto a forma de consumi-lo varia bastante de região para região e essas características serão melhor descritas e identificadas mais a diante.

## 4.2 MODELO DE ANÁLISE DE SENTIMENTOS

Nesta seção são apresentados os resultados obtidos para cada classificador de forma separada e utilizando o *ensemble voting classifier* mostrando a acurácia de seus desempenhos e o intervalo de confiança obtido pela técnica de validação cruzada. Em seguida, são apresentados os valores das métricas de validação  $R^2$ , MSE, RMSE e MAE para o melhor modelo, exibindo os resultados do obtidos pela matriz de confusão.

### 4.2.1 Validação Cruzada

Para o treinamento do modelo utilizamos a validação cruzada para garantir que o modelo construído não sobrecarregue o conjunto de treinamento. Com isso, utilizando a função `StratifiedKFold`, dividiu-se os dados em 10 *folds* em conjunto com o parâmetro `shuffle = true` a fim de garantir que os grupos serão embaralhados e que sempre haverá um percentual equivalente de dados de cada classe em cada uma das partições (tanto na de treinamento como na de teste). O principal objetivo deste procedimento é o de tornar o processo de construção de um *fold* mais amplo, minimizando o prejuízo causado pelo desbalanceamento da base de treinamento. A Tabela 3 mostra os resultados obtidos por cada classificador de forma separada.

Tabela 3 – Resultados dos classificadores. Em negrito está destacado o melhor resultado.

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

Fonte: autor.

É possível verificar que 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 em nossa base de treinamento, estes resultados estão bem próximos aos da literatura, como pode ser visto em Ankit e Saleena (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 fatores por este resultado.

#### 4.2.2 Valores das Métricas de Validação

Avaliou-se as métricas  $R^2$ , MSE, RMSE e MAE para o modelo **Ensemble Voting Classifier** que obteve o melhor resultado de acurácia média. De um modo geral, o modelo obteve também resultados bastante expressivos para essas métricas, mesmo possuindo uma base de treinamento desbalanceada como visto na subseção 3.2.2. Os resultados obtidos podem ser visualizados na Tabela 4.

Tabela 4 – 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.

Métricas	Valores
$R^2$	0.87237989
MSE	0.10578748
RMSE	0.32524987
MAE	0.08515180

**Fonte:** autor.

Um bom resultado para as métricas MSE, RMSE e MAE é dado por valores menores e quanto mais próximo de 0 melhor. Já para a métrica  $R^2$ , o modelo deve apresentar valores maiores que 0, sendo 1 o melhor valor. Os resultados foram bastante expressivos mesmo com uma base de treinamento desbalanceada. Isso se deve à utilização da técnica de validação cruzada estratificada, que permite uma randomização garantindo, dessa maneira que o conjunto de dados com classes desequilibradas possuam alguma das demais classes de menor representatividade.

#### 4.2.3 Matriz de Confusão

A matriz de confusão fornece uma visualização dos resultados sobre o quanto o modelo acertou e errou no processo de classificação. Para as classes negativa e positiva, no geral o número de acertos foi bastante significativo onde os dados que foram classificados de forma correta para estas classes representaram 70,28% no total. Devido ao desbalanceamento presente no *dataset* de treinamento, é possível verificar que os resultados para a classe neutra não foram tão satisfatórios por ser a classe de menor representatividade,

acertando apenas 0.96% das vezes. Entretanto, o aumento da representatividade dessa classe no conjunto de treinamento é dificultado pelo baixo número de *tweets* postados com esse perfil. A Figura 18 exibe os resultados obtidos na matriz de confusão.

<b>Valor real</b>	<b>Negativo</b>	2380 28.23%	472 5.60%	855 10.14%	<b>3707</b> <b>64.20%</b> <b>35.80%</b>
	<b>Neutro</b>	15 0.18%	81 0.96%	49 0.58%	<b>145</b> <b>55.86%</b> <b>44.14%</b>
	<b>Positivo</b>	450 5.34%	584 6.93%	3546 42.05%	<b>4580</b> <b>77.42%</b> <b>22.58%</b>
	<b>Soma das colunas</b>	<b>2845</b> <b>83.66%</b> <b>16.34%</b>	<b>1137</b> <b>7.12%</b> <b>92.88%</b>	<b>4450</b> <b>79.69%</b> <b>20.31%</b>	<b>8432</b> <b>71.24%</b> <b>28.76%</b>
		<b>Negativo</b>	<b>Neutro</b>	<b>Positivo</b>	<b>Soma das linhas</b>
		<b>Valor Predito</b>			

Figura 18 – Resultados da Matriz de Confusão para o conjunto de 8.432 *tweets* analisados.  
**Fonte:** autor.

É possível verificar que o modelo desenvolvido utilizando o *Voting Classifier* para a predição obteve uma acurácia de 71,24% para a classificação correta de cada classe, mostrando que mesmo com uma base de treinamento desbalanceada, foi possível obter um resultado significativo devido à utilização da técnica de validação cruzada.

Analisando o relatório de classificação da matriz de confusão, a métrica *precision* é responsável por calcular a capacidade do classificador não rotular como positiva uma amostra classificada como negativa ou outra classe, caso o problema tenha outras. Em números, para essa métrica a classe Positiva do modelo obteve o melhor resultado com cerca de 0,77 seguido da classe Negativa com 0,64 e por fim a classe Neutra com 0,56.

A métrica *recall* identifica a proporção em que está o número de verdadeiros positivos e o número de falsos negativos. Dessa forma, basicamente indica a capacidade do classificador encontrar todas as amostras de maneira correta. Para esta métrica, os valores encontrados para as classes Negativo e Positivo novamente foram melhores sendo respectivamente 0,84 e 0,80. Mais uma vez a classe Neutra obteve resultados baixos com um valor de 0,07, mostrando uma classificação bastante errônea para esta classe.

A métrica **f1-score** realiza uma média harmônica ponderada dos valores das métricas **precision** e **recall**, onde uma pontuação mais próxima de 1 possui melhor pontuação e 0 pontuações piores. Assim, mais uma vez as classes Negativa e Positiva obtiveram os melhores resultados com respectivamente 0,73 e 0,79. Quando se tem um *f1-score* baixo, tem-se que a **precision** ou a **recall** possuem um valor baixo e de fato isso ocorreu onde a classe Neutra, obteve um valor de 0,13 indicando que houve valores baixos nas outras métricas. A Tabela 5 exibe os resultados do relatório de classificação gerado pela matriz de confusão do modelo de análise de sentimentos desenvolvido.

Tabela 5 – Métricas de Avaliação para o Modelo **Voting Classifier**.

<b>Classes</b>	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
Negativo	0.64	0.84	0.73	2,845
Neutro	0.56	0.07	0.13	1,137
Positivo	0.77	0.80	0.79	4,450

**Fonte:** autor.

A análise de sentimentos dos *tweets* coletados e armazenados será mostrada após a análise de conteúdo, a fim de detalhar e explorar da melhor maneira a análise dos resultados.

### 4.3 ANÁLISE DE CONTEÚDO

Analisando o conteúdo geral ao longo das semanas, foi possível observar que em grande parte das publicações coletadas e analisadas, as categorias mais mencionadas pelos usuários do Twitter foram relacionadas aos hábitos de consumo, seguidos das características do queijo artesanal descritos na Tabela 1. A Figura 19 mostra o quantitativo de cada categoria ao longo das semanas.

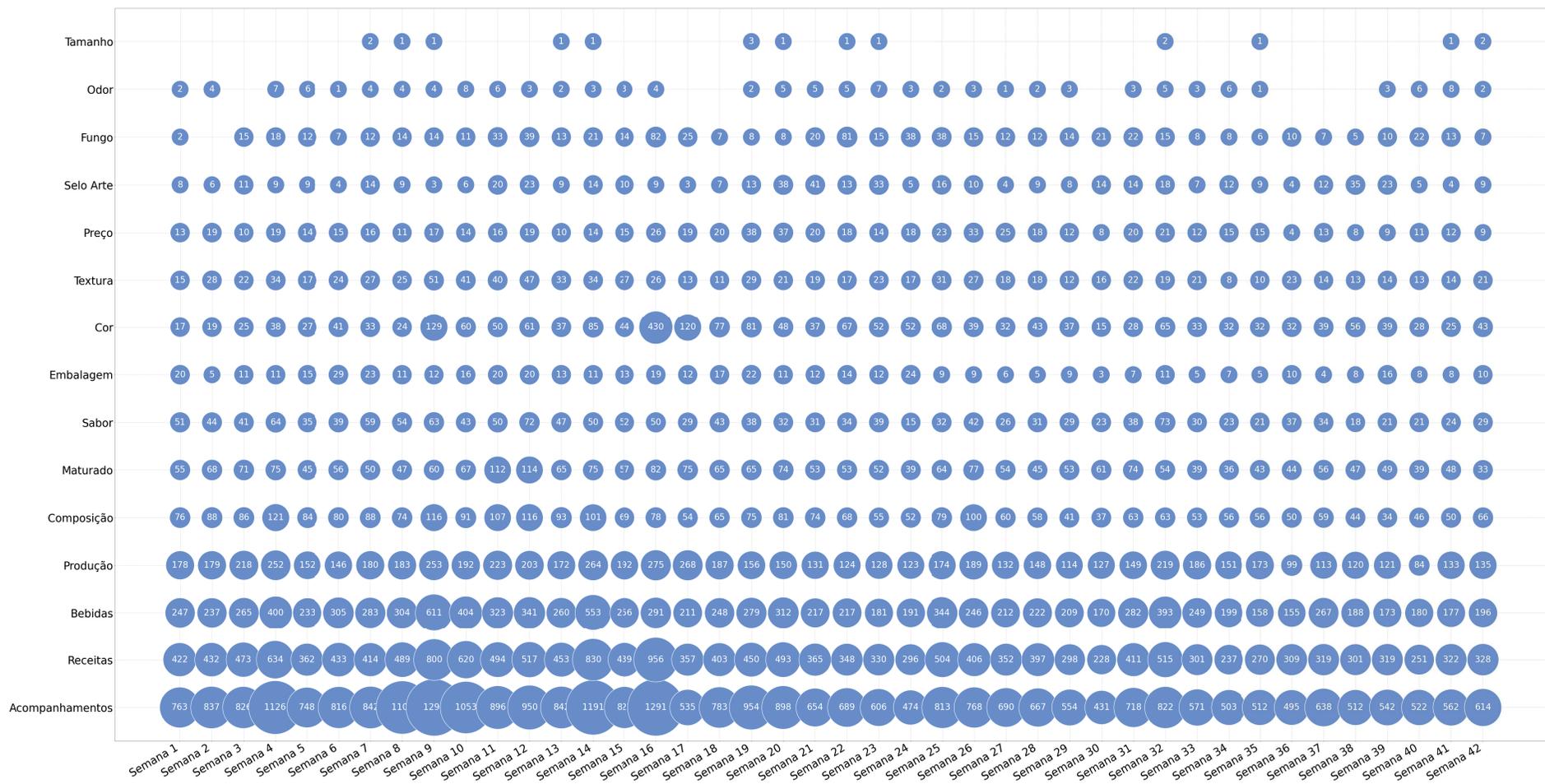


Figura 19 – Análise dos conteúdos ao longo do tempo. **Fonte:** autor.

É possível observar que ao longo das semanas, os termos que descrevem a categoria de **Acompanhamentos** foram os que obtiveram maior destaque e isso é justificado pela grande variedade de produtos que além de serem mencionados nas publicações são consumidos em conjunto com os queijos artesanais. A segunda categoria mais representativa é referente às **Receitas**, onde foi possível observar diversos pratos que possuem o queijo como ingrediente. A categoria **Bebidas**, fecha o grupo das 3 que definem os hábitos de consumo e foi possível identificar diversas bebidas associadas ao consumo dos queijos.

Das categorias que definem as características do queijo artesanal, 4 foram as de maior influência: **Produção**, **Composição**, **Maturado** e **Sabor**. A categoria **Produção** foi a que resultou em um maior número de menções ao longo das semanas, com destaque para o atributo artesanal. A segunda categoria mais influente denominada **Composição** foi capaz de identificar ao longo das semanas alguns produtos que são adicionados ao queijo artesanal em seu processo de fabricação. Já a categoria **Maturado** identificou atributos referentes ao processo de maturação dos queijos, onde verificou-se publicações mencionando queijos **maturados**, **frescos** e **curados**. E por fim a categoria **Sabor**, identificou características como **sabor suave**, **agradável** e **gosto forte** presentes em determinadas publicações.

#### 4.3.0.1 Características e Hábitos de Consumo do Queijo Artesanal

Ao realizar o processamento em cada *tweet*, foi encontrado no total cerca de 103.484 menções das características e hábitos de consumo. 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 e visualizações dos resultados.

Dessa maneira, analisando as categorias em ordem de representatividade de número de menções, a categoria **Tamanho** obteve cerca de 18 menções, não sendo tão expressiva. O atributo **queijo grande** foi o que teve um maior número de citações nesta categoria seguido de **queijo padrao** e **queijo pequeno**. Com poucas menções relacionadas ao **Tamanho**, encontrar uma caracterização geral para todos os queijos artesanais no Brasil não foi possível, entretanto, nas poucas menções coletadas, foi identificado relatos de usuários mencionando os queijos da canastra, meia cura, serro, roça, búfala e sem lactose. A Figura 20 exhibe o quantitativo de cada uma das categorias.

A categoria **Odor** foi a segunda de menor relevância ao fim das semanas de análise com cerca de 136 menções de seus atributos. O atributo **vegetal** obteve 77 menções e pôde ser observado em vários contextos. Por ser um termo bastante amplo foi identificado diversos assuntos fora da caracterização de odor, como por exemplo, usuários que se classificam como veganos ou que fazem uma dieta vegetariana descrevendo sobre o consumo

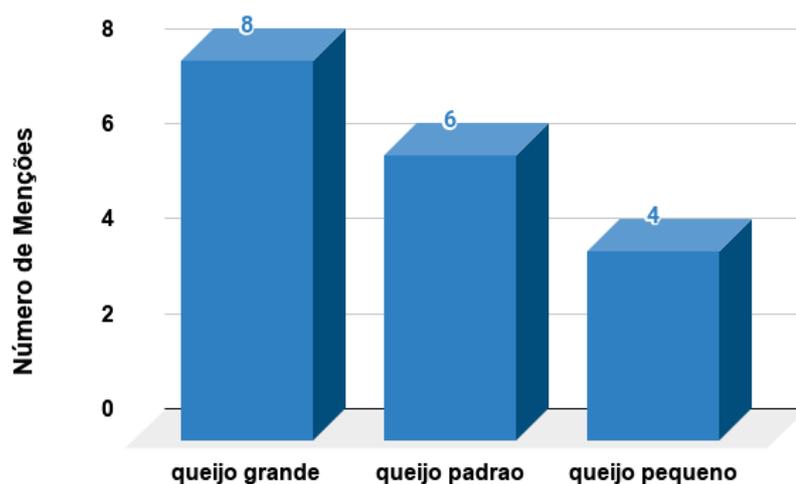


Figura 20 – Análise de conteúdo da categoria Tamanho nos *tweets* sobre queijos artesanais. **Fonte:** autor.

de produtos como leite vegetal, queijo vegetal e a presença do carvão vegetal nos queijos. O atributo **aroma** foi mencionado cerca de 31 vezes e em grande parte das vezes estava associado positivamente aos queijos com grande parte dos usuários afirmando gostar dos aromas proporcionados pelos queijos artesanais em conjunto com outros acompanhamentos.

Assim como o atributo **vegetal**, o termo **característico** esteve mais relacionado com às características que cada tipo de queijo artesanal possui devido ao regionalismo presente do que sobre o seu odor característico propriamente dito. Diversos queijos foram citados podendo ser vistos os queijos Coalho, Curado, Cabra, da Canastra, do Serro e Alagoa. Os demais atributos que definem o odor apareceram em menor número e representaram de fato publicações que estavam relacionadas ao cheiro típico do queijo, sendo tanto comentários positivos quanto negativos. A Figura 21 mostra o quantitativo de menções que cada um dos atributos obteve.

A categoria de **Fungo** foi bem mais representativa do que as já apresentadas, com 756 menções no total divididas nos atributos **mofo**, **contaminação**, **mofo azul**, **fungo branco** e **mofo natural** com respectivamente 676, 52, 24, 3 e 1 menções nos *tweets* analisados. De um modo geral, a presença do mofo traz consigo muitos questionamentos por parte dos usuários/consumidores dos queijos que apresentam tais características. Grande parte dos usuários do Twitter afirmam não ser adeptos a produtos que contenham o mofo, mas pôde ser observado que um grande número de usuários possuem a informação de que se o fungo (o mofo) for específico do queijo seu consumo não traz danos à saúde. Dessa forma, ao analisar esta categoria identifica-se usuários/consumidores que possuem na sua grande maioria um conhecimento sobre a presença ou não dos fungos nos queijos sabendo quais são seus benefícios ou malefícios advindos de seu consumo. A Figura 22 mostra em detalhes a quantidade de menções de cada um dos atributos.

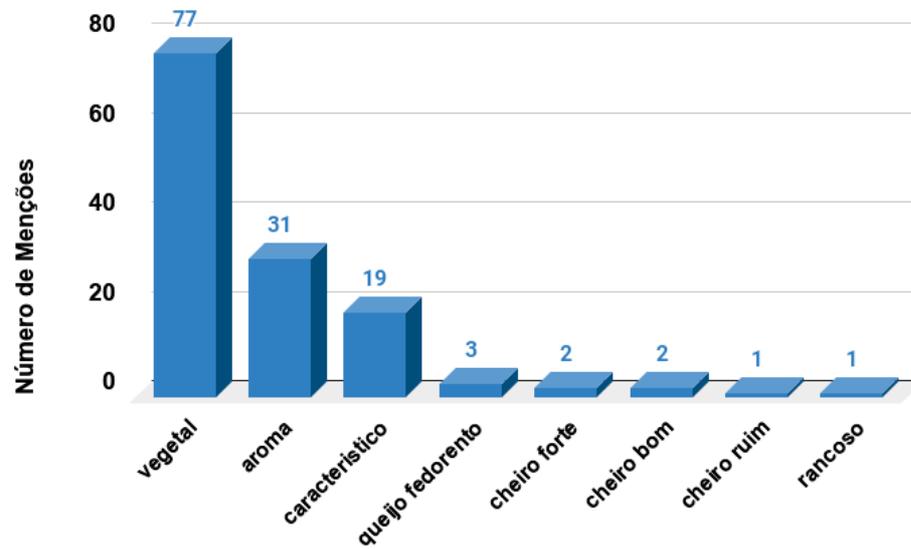


Figura 21 – Análise de conteúdo da categoria Odor nos *tweets* sobre queijos artesanais.  
Fonte: autor.

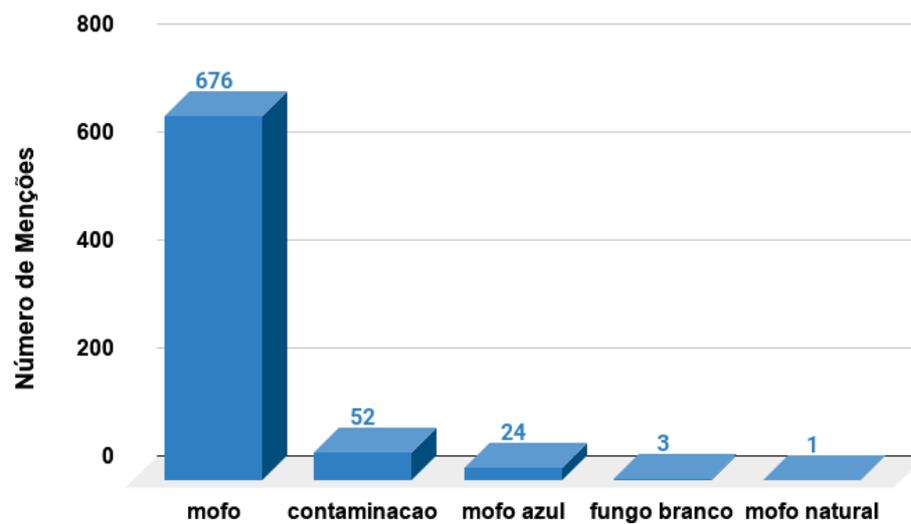


Figura 22 – Análise de conteúdo da categoria Fungo nos *tweets* sobre queijos artesanais.  
Fonte: autor.

Como dito na seção Queijo Artesanal 2.1, o selo arte pode ser considerado sinônimo de algumas características bastante importantes como qualidade, segurança e higiene sendo esses alguns dos principais requisitos que muitos usuários/consumidores buscam. Ao todo foram identificados na categoria Selo Arte 545 menções referentes aos atributos qualidade, selo arte, certificacao, higiene e inspecao com respectivamente 318, 120, 38, 35 e 34 menções. Em relação ao termo qualidade, diversos queijos foram mencionados a exemplo os queijos coalho, serrano, colonial, da canastra e do serro sendo considerados por muitos referência em qualidade, sabor e textura. O atributo selo arte, foi o segundo mais citado e grande parte das publicações foram realizadas por perfis de comunicação informando sobre queijos que conquistaram o selo arte após um ano do decreto da Lei que regulamenta o comércio de queijos, embutidos, pescados e mel em todo o território nacional do Brasil. As menções presentes nas postagens referentes aos atributos certificação, higiene e inspeção apareceram em menor número, porém, relacionaram queijos que conquistaram a certificação do selo arte por seguirem padrões de higiene e todos os protocolos de segurança necessários para produção dos queijos. A Figura 23 mostra o quantitativo de menções de cada um dos atributos desta categoria.

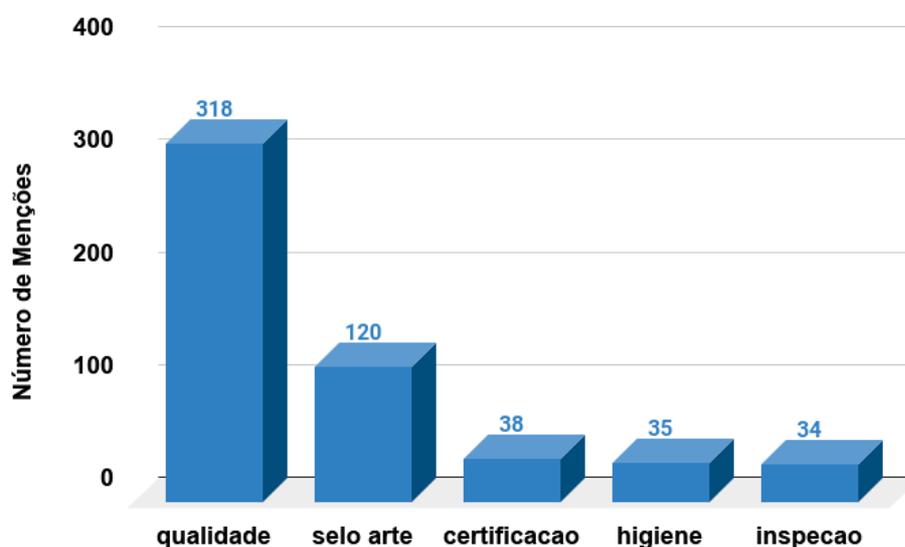


Figura 23 – Análise de conteúdo da categoria Selo Arte p nos *tweets* sobre queijos artesanais. **Fonte:** autor.

Uma categoria muito importante que também foi analisada neste trabalho foi o Preço, com 700 menções divididas nos atributos reais, queijo caro, queijo barato e preco justo. Foram contabilizados respectivamente 665, 21, 7 e 7 menções a cada atributo. O atributo reais foi bastante mencionado porém em diversos contextos, podendo estar associado aos próprios queijos artesanais, e a receitas e lanches que possuem o queijo como um dos ingredientes. Ao analisar o preço de um modo geral, há relatos de queijos sendo vendidos por R\$ 50,00 reais como é o caso do queijo da canastra e a R\$ 60,00 reais

o queijo coalho no estado do Pernambuco. A Figura 24 exhibe a quantidade de menções de cada um dos atributos presentes nesta categoria.

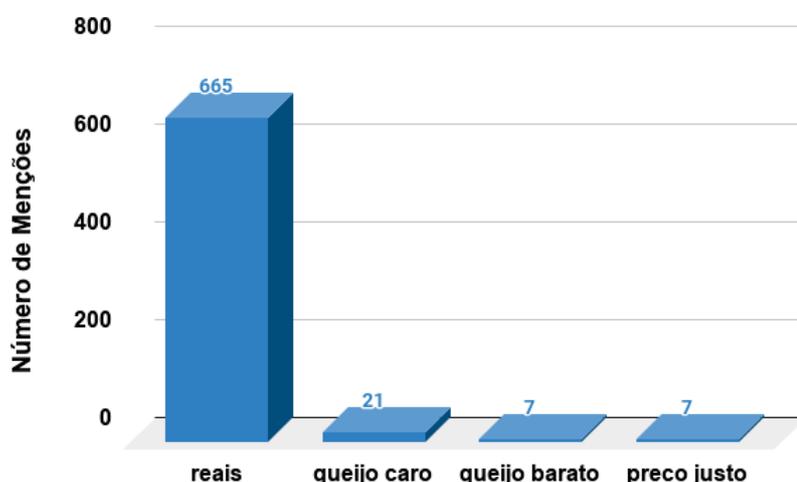


Figura 24 – Análise de conteúdo da categoria Preço nos *tweets* sobre queijos artesanais. Fonte: autor.

A categoria Textura obteve no total 985 menções referentes aos atributos cremoso, seco, mole, duro, firme, macio, gorduroso, amanteigada, olhaduras, fechada, umido, casca dura, pastoso, pegajoso, massa macia, elastica, oleoso e untuoso, que definem essa categoria. De todos os atributos mencionados anteriormente, cremoso e seco foram os que obtiveram maiores números de menções com 279 e 273, respectivamente e foi possível verificar que ambos se referenciavam à textura dos queijos. Os queijos Canastra, Mogiana, Serro, Coalho e Curados foram bastante mencionados sendo possível identificar várias características físicas e reais dos queijos sendo publicadas. A Figura 25 mostra o quantitativo de menções ao longo tempo.

Analisando a categoria Cor, foi contabilizado um total de 2.389 menções. O atributo queijo azul foi o que teve o maior quantitativo de menções (1.265) seguido de branco com 905 menções, amarelo com 190, pintado com 11, amarelo claro com 11, amarelado com 5 e por fim amarelo escuro com 2 menções. O grande destaque para o atributo queijo azul se dá pelo fato dessa característica ser muito marcante neste queijo sendo um dos principais diferenciais para os demais. As cores branco e amarelo estavam muito associadas aos queijos Coalho, Tradicional e Curado. Já a cor amarelo claro esteve bem relacionada com o queijo Tulha. A Figura 26 mostra o total de menções por atributo.

A categoria Embalagem, apresentou no total 517 menções, sendo 195 referentes ao atributo palha, 127 à embalagem, 62 à papel, 56 à plástico, 50 à caixinha, 14 à vácuo e 13 à cestas. O termo palha estava mais associado ao produto batata palha, estando presente em correntes de mensagens relacionadas com os pratos favoritos, aparecendo em conjunto com o *stroganoff* o que inviabilizou uma análise mais assertiva para esse atributo.

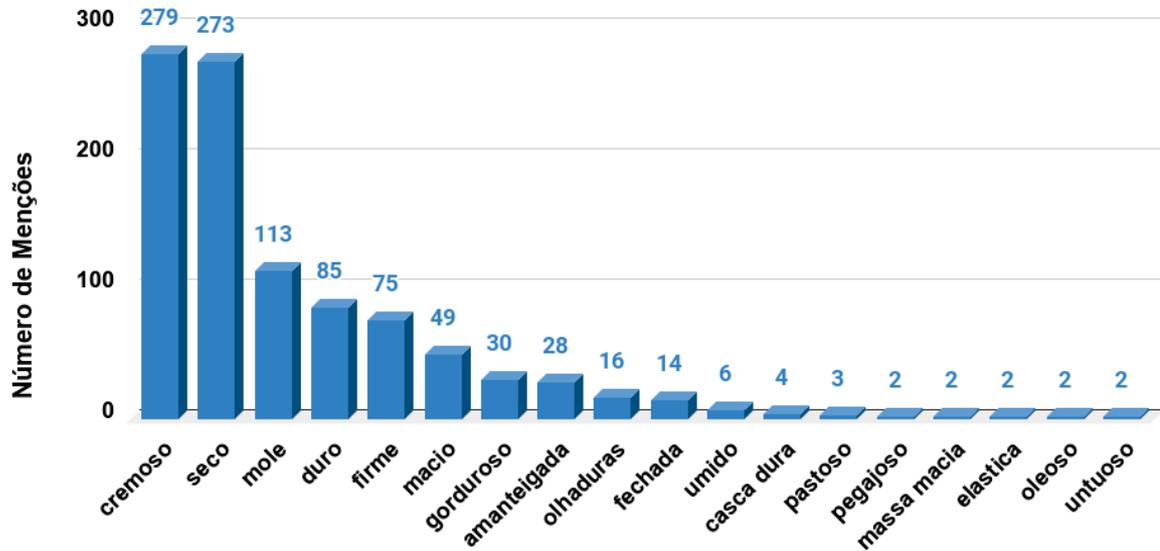


Figura 25 – Análise de conteúdo da categoria Textura nos *tweets* sobre queijos artesanais.  
**Fonte:** autor.

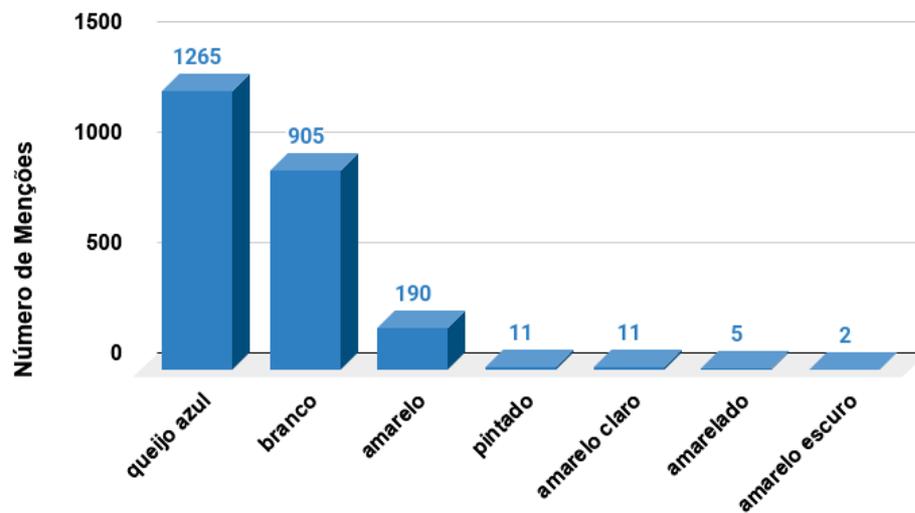


Figura 26 – Análise de conteúdo da categoria Cor nos *tweets* sobre queijos artesanais.  
**Fonte:** autor.

Os demais atributos estavam relacionados tanto ao contexto das embalagens dos queijos quanto a de produtos como hambúrgueres artesanais ou outros que possuem os queijos como ingredientes. A Figura 27 mostra o gráfico da representatividade de menções de cada um dos atributos que compõem a categoria **Embalagem**.

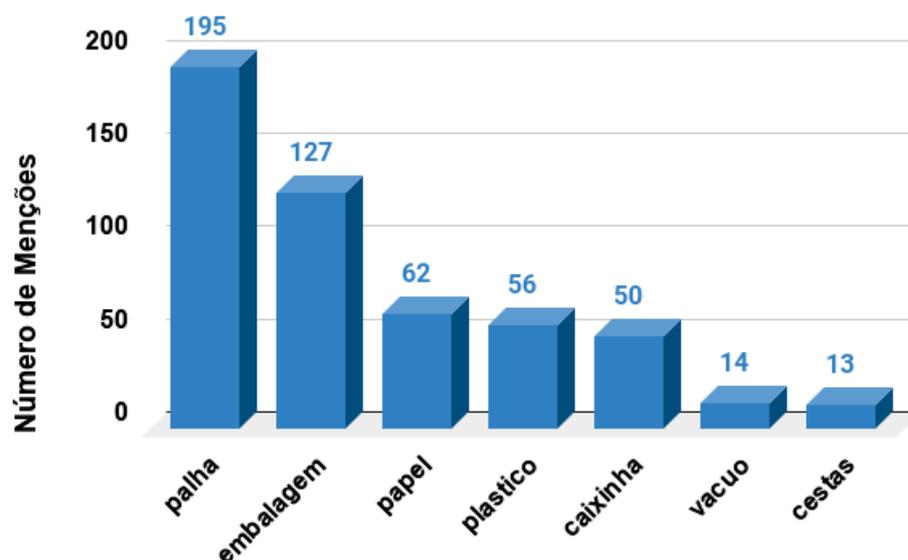


Figura 27 – Análise de conteúdo da categoria **Embalagem** nos *tweets* sobre queijos artesanais. **Fonte:** autor.

Uma das categorias mais importantes em um queijo é o seu sabor. Dito isso, a categoria **Sabor** obteve resultados bastante interessantes e ao todo cerca de 1.769 menções puderam ser encontradas, divididas em 17 atributos sendo **sal**, **salgado**, **azedo**, **saudavel**, **amargo** com respectivamente 738, 336, 197, 150, 107 menções. Com relação ao **sal**, muitos usuários relataram receitas que continham o **sal** como ingrediente. Já o termo **salgado** estava muito associado a queijos que continham um alto teor de **sal**, além de estar relacionados a lanches que tinham também os queijos. O termo **azedo** esteve bastante presente e em grande parte de suas menções associava-se ao **polvilho azedo** para a confecção de pães de queijo artesanal. O termo **saudavel** esteve muito relacionado a consumidores preocupados com uma alimentação mais regrada, saladas com algum queijo artesanal e até usuários se indagando sobre a saudabilidade de certos queijos. As menções ao termo **amargo**, estavam bastante relacionadas com alguns acompanhamentos do queijo na hora do consumo. Um exemplo é a combinação de queijo **Coalho** assado com chocolate amargo que foi mencionado por um usuário. As demais características apareceram em menor número, mas estavam bastante presentes, referindo-se às características mais marcantes de diversos queijos. A Figura 28 mostra em detalhes o quantitativo de menções de cada atributo.

Analisando a categoria **Maturado**, ao todo foram encontradas 2.653 menções sendo 1.760 referentes ao termo **curado**, 561 ao termo **fresco** e 332 ao termo **maturado**. O

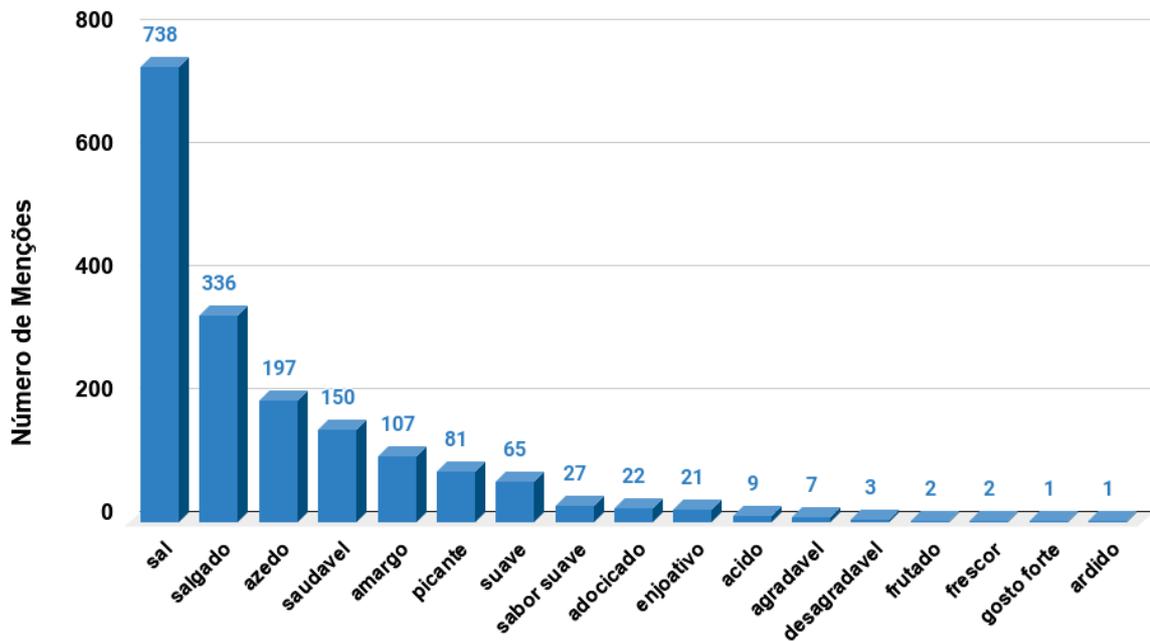


Figura 28 – Análise de conteúdo da categoria Sabor nos *tweets* sobre queijos artesanais. **Fonte:** autor.

termo curado esteve bastante presente em *posts* com opiniões e também com brincadeiras associando o termo curado com a cura da doença causa pelo novo coronavírus. Já o atributo *fresco*, esteve diretamente relacionado aos queijos frescos, assim como o termo *maturado*. Por estar analisando queijos artesanais que na sua grande maioria são curados/maturados têm-se essa maior preferência por queijos que possuem essas características. A Figura 29 exhibe a quantidade de cada um desses atributos ao longo do tempo de estudo em questão.

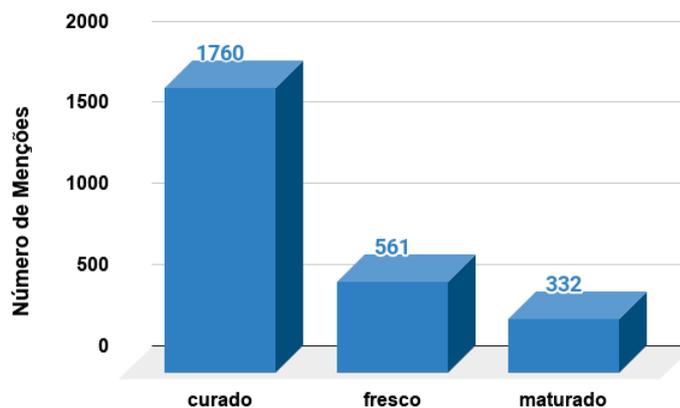


Figura 29 – Análise de conteúdo da categoria Maturado nos *tweets* sobre queijos artesanais. **Fonte:** autor.

A categoria Composição, teve no total 3.290 menções com 17 atributos sendo a quinta categoria mais representativa englobando atributos como cebola, oregano, manjericao, puro, temperado, tomate seco, defumado e beterraba, que obtiveram

respectivamente 1.011, 845, 248, 215, 184, 160 e 112 citações em publicações. A cebola, foi um produto que na grande maioria de suas aparições esteve relacionado com o queijo através da presença em pratos, lanches e receitas, ou seja, não fazendo parte do próprio queijo em si. O atributo *oregano*, esteve muito associado ao queijo coalho, podendo ser considerado uma característica bem marcante de quem consome este produto. Aparecendo em menor número, mas ainda bastante significativo, o *manjericao* assim como o *orégano*, esteve relacionado como um acompanhamento em determinadas receitas. Analisando ao termo *puro*, foi possível verificar muitos comentários de usuários mencionando o consumo do queijo artesanal sem o acréscimo de nenhum outro acompanhamento. Os termos *temperado*, *tomate seco*, *defumado* e *beterraba* foram adicionados nessa categoria devido a muitos queijos possuírem tais características, porém, ao fazer a análise destes termos, o conteúdo publicado está mais relacionado com acompanhamentos de determinadas receitas que os utilizam do que propriamente se referindo ao queijo em si. Os demais atributos que obtiveram menores números de menções ocorreu o inverso. Grande parte dos *tweets* se referiam a sua presença como parte do queijo artesanal. A Figura 30 mostra o número de menções de cada atributo analisado nesta categoria.

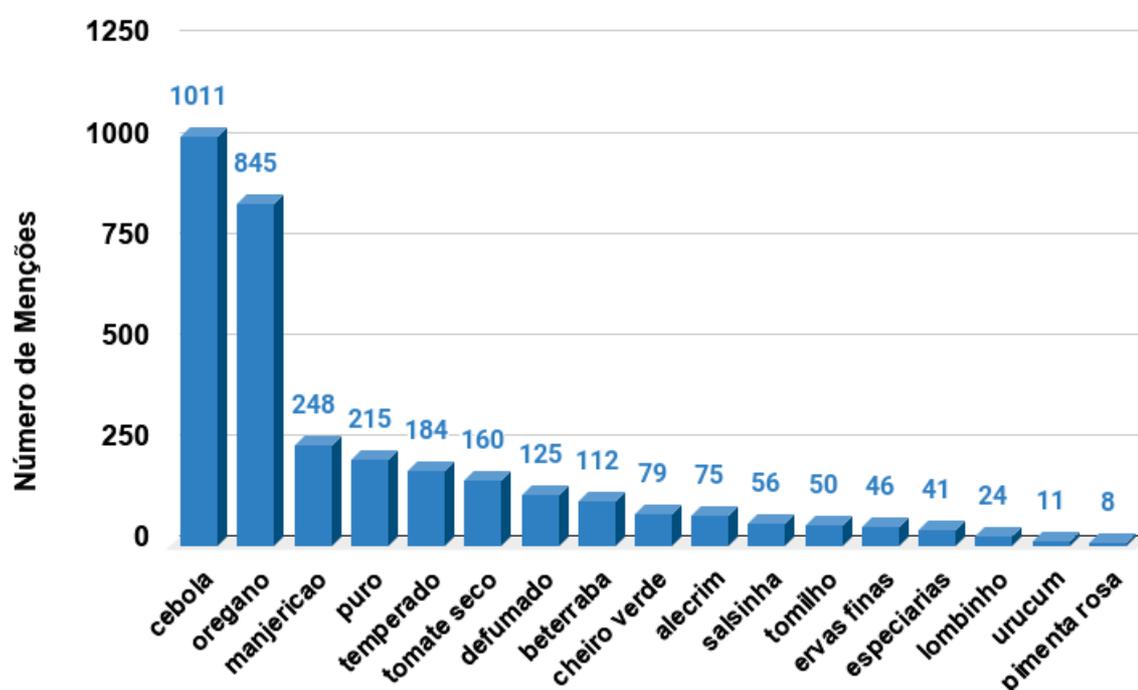


Figura 30 – Análise de conteúdo da categoria Composição nos *tweets* sobre queijos artesanais. **Fonte:** autor.

A categoria Produção obteve números bastante expressivos e no total 7.232 menções foram identificadas. O atributo *artesanal* foi o que mais foi citado com cerca de 4.502 aparições, seguido de *tradicional* com 2.448 citações, *natural* com 250 e *familiar* com 32. O termo *artesanal* obteve um número bastante expressivo e foi possível observar

muitos usuários expressando opiniões sobre os queijos artesanais, mas também produtos feitos artesanalmente, sendo possível identificar menções a hambúrguer, cerveja e pizzas produzidas de maneira artesanal. Com relação ao atributo **tradicional**, as publicações que mencionaram tal característica variaram bastante de contexto. Umas estavam relacionadas a sabores tradicionais de determinados produtos, outras referiam-se a procedimentos tradicionais para a confecção de produtos. O termo **natural** esteve muito associado com o tipo de leite usado para a produção dos queijos, ou referindo-se a sucos naturais e a pães artesanais de fermentação natural. O atributo **familiar**, foi muito relacionado a publicações de usuários mencionando práticas familiares presentes no modo de produção dos queijos artesanais juntamente a publicações de notícias sobre alguns governos estaduais, entre eles o de Minas Gerais reconhecendo as regiões de Alagoa e Mantiqueira como regiões produtoras de queijos artesanais. A Figura 31 mostra os números de menções para cada atributo.

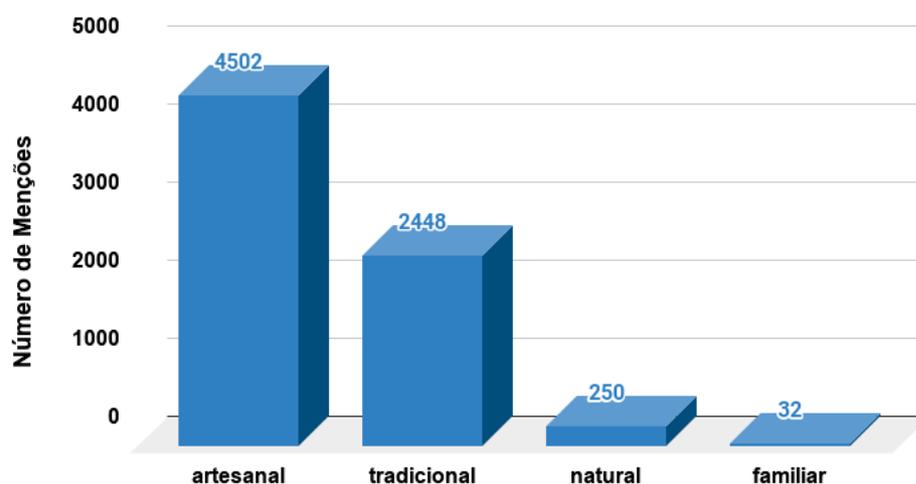


Figura 31 – Análise de conteúdo da categoria **Produção** nos *tweets* sobre queijos artesanais.  
**Fonte:** autor.

Como dito anteriormente, os resultados referentes aos hábitos de consumo foram os mais expressivos vistos nesta análise de conteúdo. Com isso, a categoria **Bebidas** obteve no total 12.637 menções, sendo os atributos **cafe**, **leite**, **suco**, **vinho**, **cerveja** e **cha** com os de maiores números de citações nas publicações com respectivamente 3.523, 3.191, 1.969, 984, 943, 823 menções. Como era de se esperar o **cafe** foi a bebida de maior influência e é nítido o quanto é amplamente difundido em todo o país. O **leite**, foi a segunda bebida que mais foi mencionada e houve momentos que estava associada ao leite cru utilizado para a produção dos queijos artesanais e em outros combinada com o café, sendo estes um dos hábitos de consumo mais marcantes associados ao queijo. O atributo **suco** obteve também uma representatividade bastante expressiva sendo muito citado em correntes de sabores favoritos feitas pelos usuários.

O termo *vinho* possui uma forte relação de harmonização com queijos artesanais e o consumo de ambos produtos pôde ser bastante observado nas publicações. A *cerveja* é uma das bebidas mais consumidas em todo o Brasil e foi muito mencionada em conjunto com o queijo *Coalho*, muito usado em churrascos em determinadas regiões do Brasil com destaque para a região Sudeste com 81,7% das menções a este termo. As aparições do atributo *chá* estiveram presentes em muitos *tweets* e assim como o termo *suco*, apareceram em muitas correntes de mensagens sendo uma delas referente ao “Nível de Frescura” onde os usuários respondiam se gostavam ou não de determinados produtos. Os demais atributos aparecem em menor número, porém, mostram o quão diversificado é o consumo dos queijos artesanais. A Figura 32 mostra os resultados que foram encontrados para esta categoria.

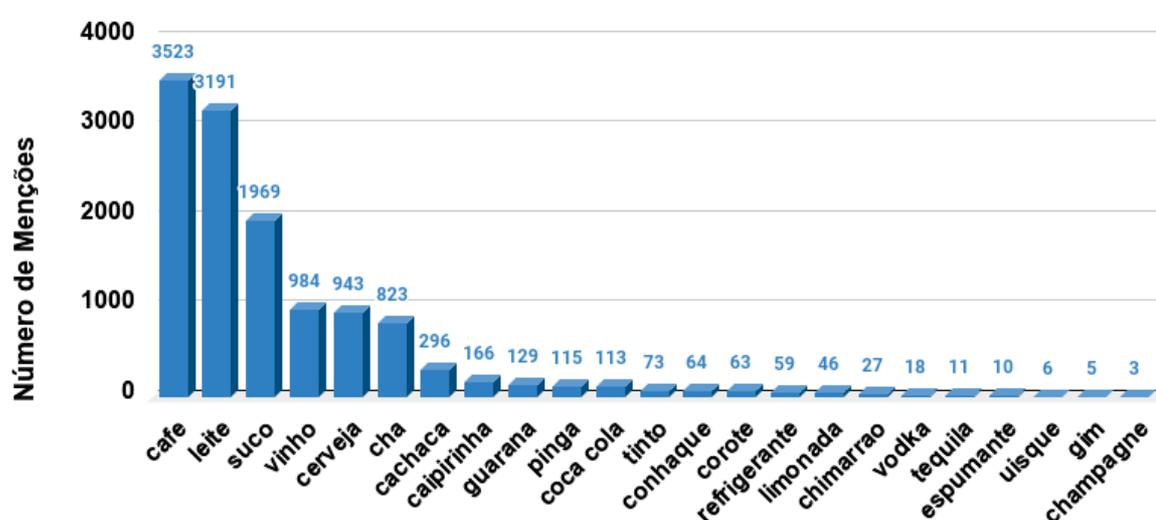


Figura 32 – Análise de conteúdo da categoria Bebidas nos *tweets* sobre queijos artesanais. Fonte: autor.

A categoria de receitas foi a segunda de maior representatividade e foram identificadas no total 20.739 menções às receitas utilizadas para análise neste estudo. Seis atributos obtiveram mais de 1.000 menções às receitas de interesse que foram *pão queijo*, *tapioca*, *pizza*, *hambúrguer*, *carne sol* e *arroz*. Analisando o atributo *pão queijo*, foi observado que duas possíveis situações poderiam ocorrer, onde a primeira relacionada a *pão de queijo* e a segunda a *pão com queijo*. Isso ocorre devido à etapa de remoção das *stopwords*. A *tapioca* é uma comida tipicamente nordestina e foi bastante citada nos *tweets* muito relacionados com o queijo *Coalho*. Por mais que este produto por si só esteja relacionado a saudabilidade, por apresentar baixos índices calóricos além de auxiliar na perda de peso (GORETTI, 2021), seu consumo no universo dos queijos mostra um caráter tanto indulgente quanto saudável. Quando os acompanhamentos são por exemplo o doce de leite, a goiabada, o leite condensado ou o mel este consumo é classificado como indulgente e quando relacionam acompanhamentos como tucumã, banana, coco, chia, ovos mostram pessoas preocupadas com uma alimentação mais saudável.

A **pizza** e o **hamburguer** foram os lanches que mais se destacaram com a presença dos queijos. O termo **carne sol**, refere-se ao prato tipicamente também nordestino carne de sol e em grande parte de suas menções esteve relacionado com o queijo coalho. O termo **arroz**, foi muito presente também em publicações que envolviam um outro prato típico da região nordeste do Brasil que é o “Baião de Dois” sendo novamente o queijo coalho o mais citado. De um modo geral, essas análises mostram que o fato do queijo Coalho ter sido o queijo artesanal mais citado, muitas das receitas típicas da região Nordeste do Brasil puderam ser observadas. As demais receitas foram publicadas em menor número, porém em todos a presença do queijo é muito marcante mostrando como seu consumo é diversificado. A Figura 33 mostra a quantidade de menções que as receitas analisadas neste trabalho tiveram.

Por fim, a categoria que obteve o maior número de menções é referente aos acompanhamentos que são consumidos em conjunto com os queijos. Ao todo, 40.459 menções foram identificadas. Diversos acompanhamentos puderam ser extraídos das publicações sendo que os atributos **pao** com 6.094 menções, **carne** com 3.109 menções, **bolo** com 2.401 menções, **banana** com 1.950 menções, **frango** com 1.850 menções, **assado** com 1.799 menções, **mel** com 1.652 menções, **batata** com 1.608 menções, **bacon** com 1.390 menções, **doce** com 1.361 menções, **linguica** com 1.252 menções, **pamonha** com 1.058 menções e **pimenta** com 1.049 menções foram os 13 acompanhamentos que tiveram mais de 1.000 menções. O termo **pao** esteve muito associado a alho, linguica, queijo e doce, sendo estas combinações muito comuns e bastante populares em todo o Brasil. O termo **carne** e **bolo**, foram o segundo e o terceiro a serem mais mencionados pelos usuários/consumidores e ambos foram citados em diversas correntes podendo também verificar publicações referentes à churrascos, café da tarde com a presença de bolo e queijo, dentre outros.

O termo **banana** esteve relacionado com o queijo coalho, sendo mencionada na farofa, na cartola - outra receita típica do Nordeste do Brasil, podendo ser consumida assada ou frita -, com canela e também com cuscuz. O termo **assado** estava associado ao modo de consumir o queijo coalho e foi mencionado muitas vezes, relacionado com os churrasquinhos vendidos em praias brasileiras. Os termos **batata** e **bacon**, em grande parte das vezes foram mencionados juntos por estarem presentes em porções, lanches e hambúrgueres. Os atributos **linguica** e **pimenta** foram descritos em diversos contextos como receitas, em divulgação para vendas, no consumo de dadinhos de queijo com pimenta, dentre outros.

O atributo **doce** esteve em muitas publicações mencionadas com os queijos e foi possível verificar a presença dos doces de leite, abóbora, figo e coco, goiabada e a própria pamonha, tanto a doce quanto a salgada. Já o termo **mel**, esteve muito relacionado à tapioca, em conjunto novamente com o queijo coalho. Os demais atributos de acompanhamentos dos queijos artesanais foram bastante diversificados e a Figura 34

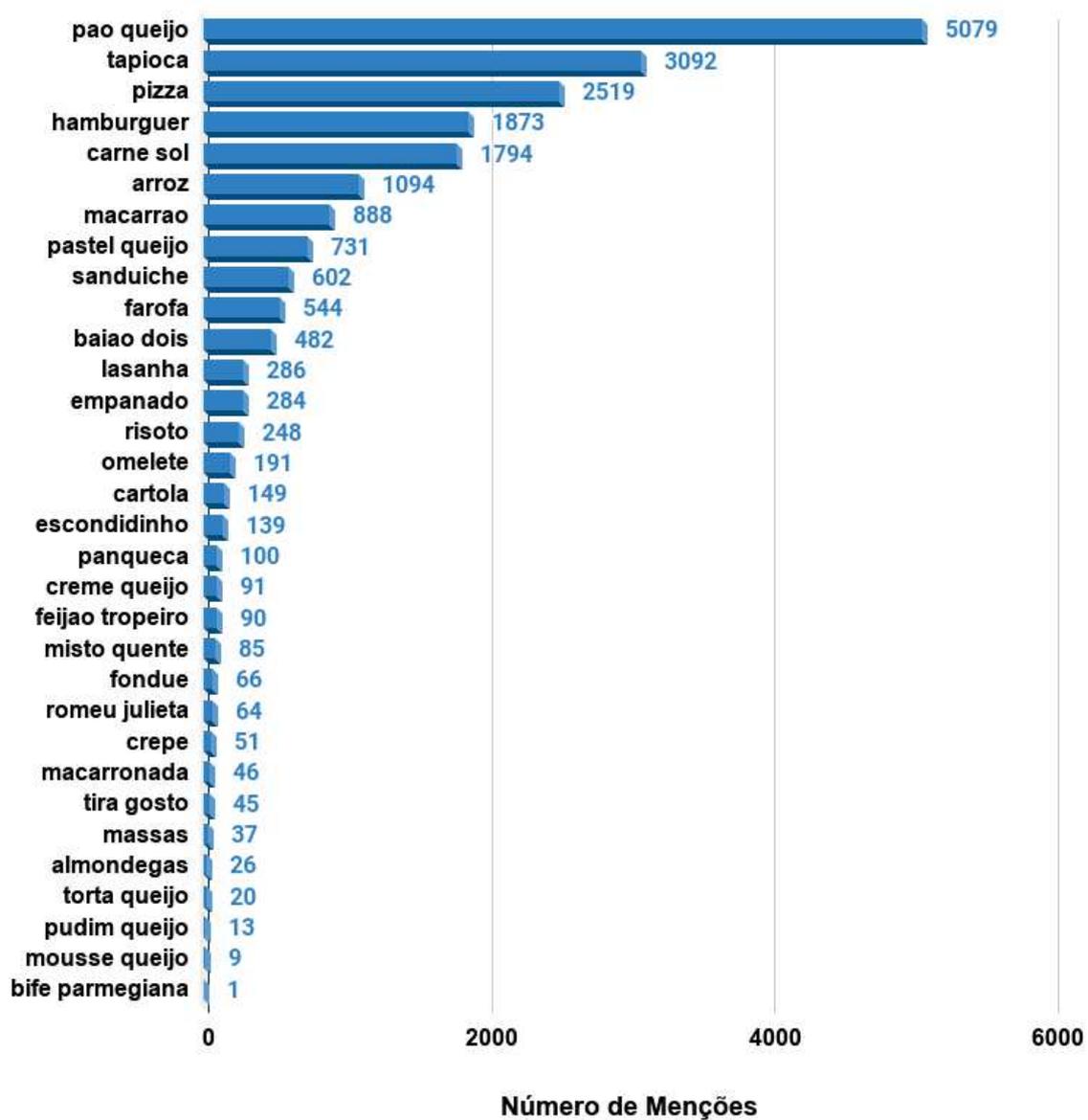


Figura 33 – Análise de conteúdo da categoria Receitas nos *tweets* sobre queijos artesanais.  
 Fonte: autor.

mostra a visão geral da quantidade de menções de cada atributo.

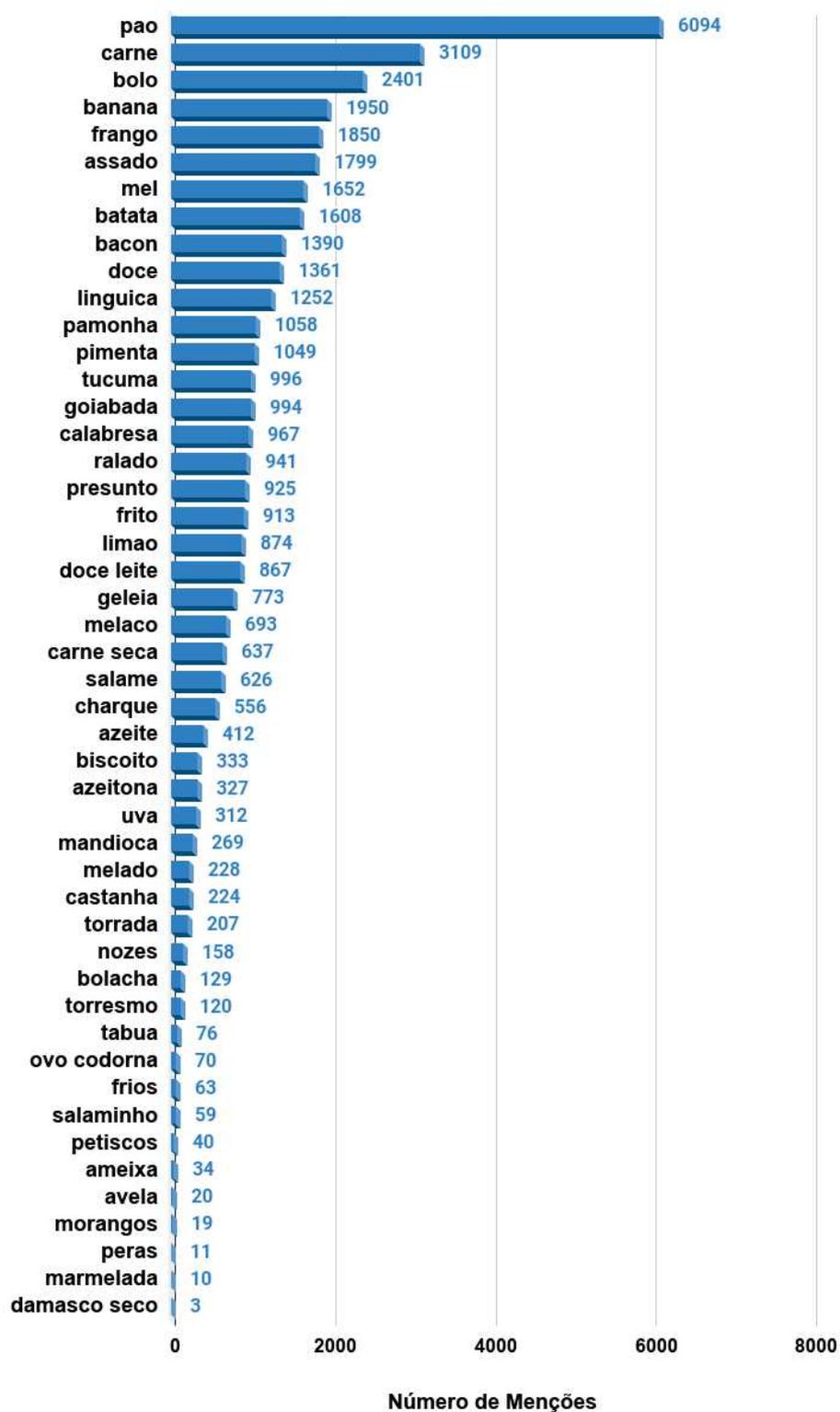


Figura 34 – Análise de conteúdo da categoria Acompanhamentos nos *tweets* sobre queijos artesanais. Fonte: autor.

#### 4.3.1 Análise da Polaridade dos Sentimentos e Consumo ao Longo das Semanas

Através do modelo de análise de sentimento desenvolvido, a fim de obtermos uma visão geral da polaridade do sentimento presente nos *tweets* publicados pelos usuários, a Figura 35 exibe o quantitativo total de classificações negativas, neutras e positivas.

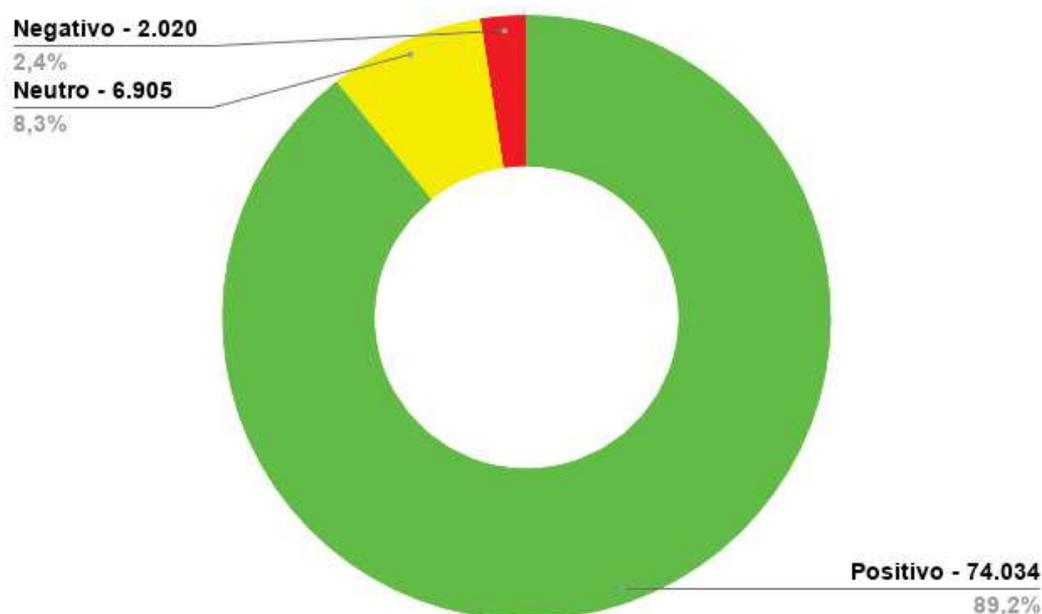


Figura 35 – Análise geral da polaridade dos sentimentos dos *tweets* sobre queijos artesanais ao longo das semanas de estudo (30/04/2020 a 18/02/2020). **Fonte:** autor.

De um modo geral, o sentimento positivo foi o que mais se destacou com 89,2% de classificações e isso está diretamente relacionado com o contexto em que os usuários publicaram os *tweets*. 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%. Sem características de destaque presentes nas frases, a predição da polaridade do sentimento realizada pelo modelo torna-se inconclusiva, mostrando neutralidade na publicação.

As classificações negativas apareceram em menor número representando um total de 2,4% do total. As classificações negativas no contexto dos queijos artesanais se deram muito relacionadas a pessoas intolerantes à lactose. Também foi possível notar nas classificações negativas, afirmações sobre os incômodos proporcionados pelo mofo presente nos queijos. Analisando semana a semana a proporção dessas classificações, pode-se ver na Figura 36 como ficou este panorama.

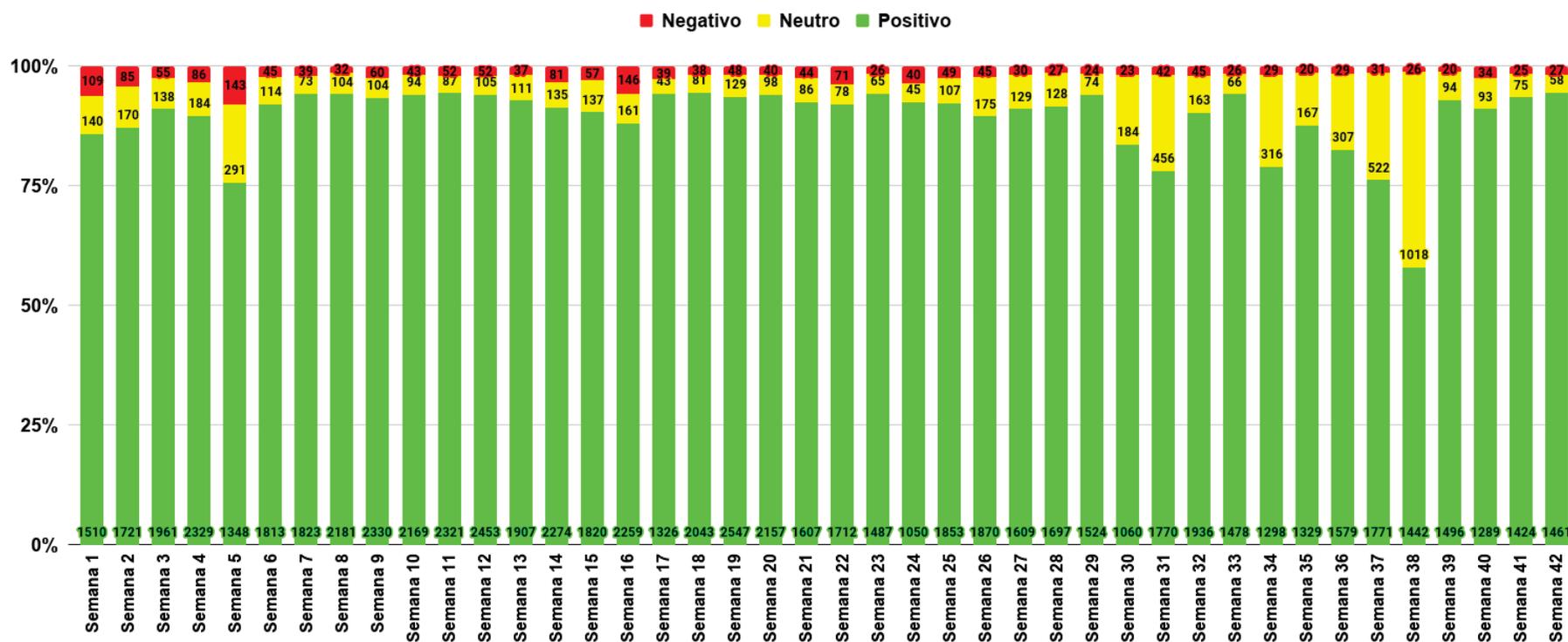


Figura 36 – Análise da polaridade dos sentimentos ao longo das semanas de estudo (30/04/2020 a 18/02/2020). Fonte: autor.

Analisando semanalmente a proporção das classificações dos *tweets*, é possível observar que a polaridade do sentimento positivo foi predominante sobre os sentimentos neutros e negativos. Em algumas semanas, foi possível observar alguns acontecimentos que impactaram o número de *tweets* e conseqüentemente a polaridade dos sentimentos dos mesmos. Na semana 4 foi identificado um pico de 2.599 *tweets* que estão diretamente relacionados com a aparição dos queijos artesanais em um concurso de culinária na televisão com alta repercussão nacional chamado “Mestre do Sabor”. Os *chefs* de cozinha tinham o desafio de incluir alguns dos queijos artesanais especificados pelo programa em seus pratos para seguirem na disputa. Com isso, a Figura 37 exibe as palavras mais frequentes presentes nos *tweets* desta semana.

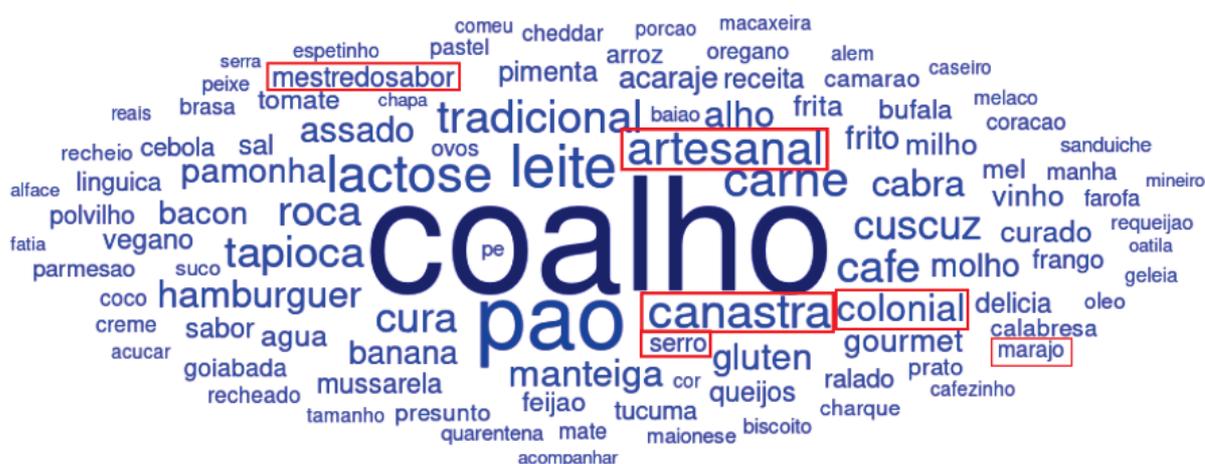


Figura 37 – Nuvem de palavras referente à semana 4 (21-05-2020 a 28-05-2020). **Fonte:** autor.

É possível observar a aparição de termos como *coalho* sendo o de maior representatividade e por esse motivo não foi destacado em vermelho como os termos *artesanal*, *mestredosabor*, *colonial*, *canastra*, *serro* e *marajo* sendo os 4 últimos referentes a alguns queijos que foram utilizados no desafio do programa de culinária citado. Outros termos também apareceram com bastante representatividade a exemplo do *pão* podendo este estar relacionado a pão com queijo, pão de queijo ou até mesmo pão de alho. O termo *carne* também citado, é referente à combinação por exemplo de carne seca com queijo *coalho*. Nas bebidas, os termos *leite*, *cafe* e *vinho* foram os de maior destaque.

Os termos *tapioca* e *cuscuz* são comidas tipicamente nordestinas e foram bastante citadas nos *tweets* muito relacionados com o queijo *Coalho*. Por mais que estes dois produtos por si só estejam relacionados à saudabilidade, seu consumo no universo dos queijos mostra um caráter tanto indulgente quanto saudável. Dessa maneira, na semana 4 é possível ter a percepção de que há uma certa tendência de se consumir os queijos artesanais com produtos indulgentes, pois pela análise de palavras mais frequentes presente na nuvem de palavras, houve uma grande aparição de produtos considerados indulgentes

que são aqueles capazes de fornecer sensação de prazer e bem estar em momentos de aflição.

Passados quase 2 meses de coleta de dados, na semana 9 ocorreu outro pico com um total de 2.494 *tweets*, sendo a polaridade positiva novamente predominante sobre as classes neutra e negativa. Devido ao isolamento social causado pela pandemia, não foi possível a realização das tradicionais festas juninas, o que levou aos usuários do Twitter criar correntes relacionadas à “Comidas de Festa Junina” e “Sabores Favoritos” além de publicações referentes aos próprios queijos. A Figura 38, mostra as palavras de maior destaque nesta semana de análise.



Figura 38 – Nuvem de palavras referente à semana 9 (25-06-2020 a 2-07-2020). Destacado em vermelho os queijos artesanais mais mencionados nesta semana, em verde as palavras mais frequentes relacionadas à corrente *Comidas de Festa Junina* e em ciano as palavras mais frequentes relacionadas à corrente *Sabores favoritos*. **Fonte:** autor.

É possível observar uma grande mudança nas publicações, e os termos de maior destaque foram os que estavam nas duas correntes, onde a primeira é referente a *Comidas de Festa Junina* e a outra *Sabores Favoritos* com 56 e 241 *tweets* publicados, respectivamente. Diversos termos obtiveram um maior destaque a exemplo de *pipoca*, *canjica*, *pamonha*, *comidas*, *junina* e *pacoca* que são referentes à primeira corrente. Já com relação à segunda corrente, os termos *macarrao*, *sopa*, *torta*, *esfiha*, *carne*, *sopa*, *sabores* e *favoritos* foram os mais evidenciados. Os queijos artesanais também apareceram sendo o queijo *coalho* o de maior representatividade e por isso não foi destacado de vermelho, o da *canastra*, o *colonial* e o termo *artesanal* em menor destaque, sendo estes os mais mencionados pelos usuários.



#### 4.3.2 Análise de Consumidores Reais e Potenciais Consumidores dos Queijos Artesanais

Utilizando os termos presentes na Tabela 2, foi realizada a mineração em todos os *tweets* a fim de identificar nestes dados informações referentes ao ato de consumir e a vontade de consumir os queijos em conjunto com seus acompanhamentos e receitas. Com isso a Figura 44, mostra os resultados obtidos ao longo das semanas.

No total foi possível contabilizar cerca de 29.684 menções aos termos referentes às categorias **Consumidores reais** e **Potenciais consumidores** referentes ao ato de consumir queijos artesanais em conjunto com seus acompanhamentos e receitas, sendo 22.064 menções pertencentes aos termos da primeira categoria e 7.620 menções aos termos da segunda. É possível verificar que de certo modo, os usuários/consumidores que utilizam a rede social do Twitter em média, 74,3% das vezes ao longo das semanas disseram consumir queijos artesanais ou receitas que o tenham como um dos ingredientes, publicando comentários de polaridade positiva na grande maioria das vezes como descrito nas Figuras 35 e 36.

De modo parecido, foi possível identificar que aproximadamente 25,7% dos usuários/consumidores que mencionaram algum dos termos referentes à **Potenciais consumidores** de consumir em seus *tweets*, não consumiram de fato, porém associaram o consumo dos queijos artesanais e ou receitas que o tenham como ingredientes, de forma positiva. Quando esteve associado de forma negativa, no geral foi possível identificar comentários associados a impossibilidade momentânea de consumir determinados queijos artesanais ou produtos que o tenham como ingrediente.

Esses resultados são bastante interessantes, pois mostram o potencial que a ferramenta do Observatório do Consumidor possui para a realização de extração de informações sobre o consumo presente em dados das redes sociais. Com o uso desta ferramenta, pesquisas de mercado que poderiam levar semanas e até meses para serem aplicadas, avaliadas e analisadas, desde que se tenha um conjunto de palavras-chave bem definidas, dados históricos e correntes, em pouco tempo torna-se possível identificar diversas respostas.

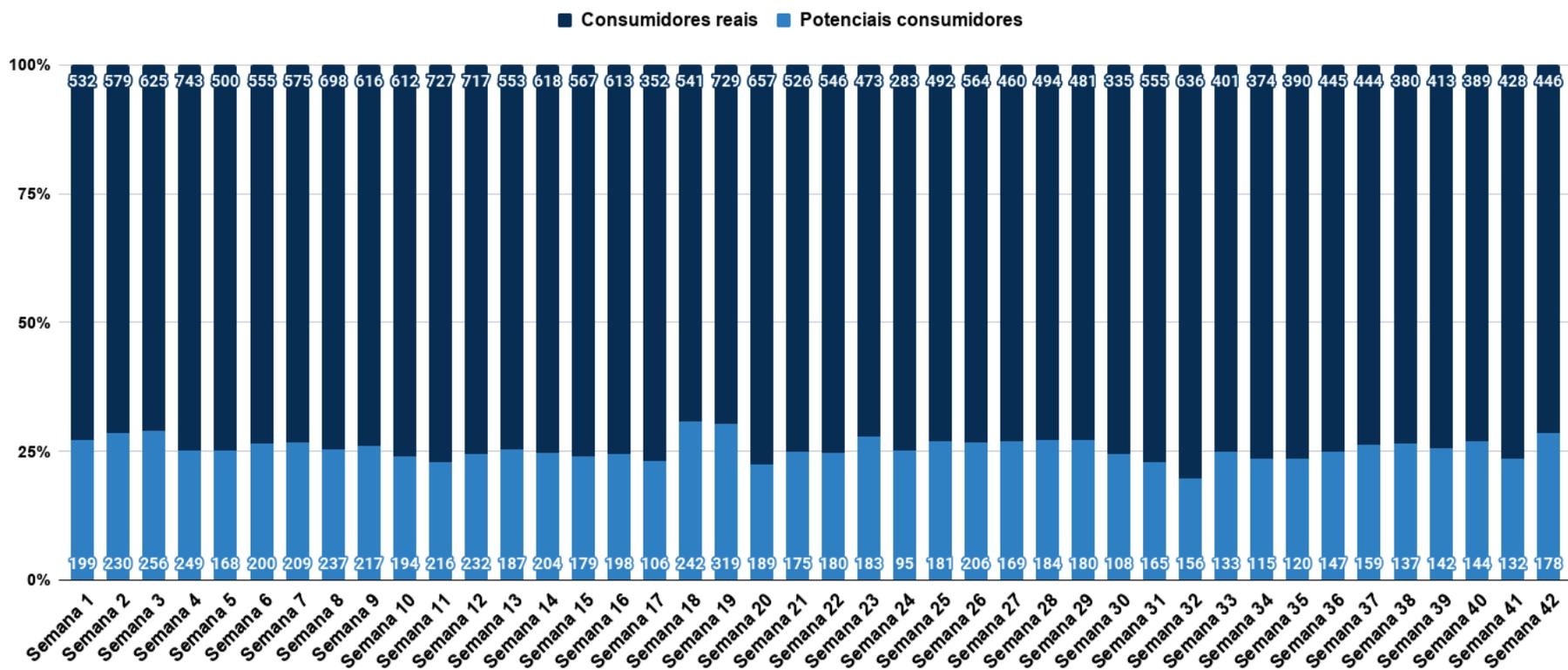


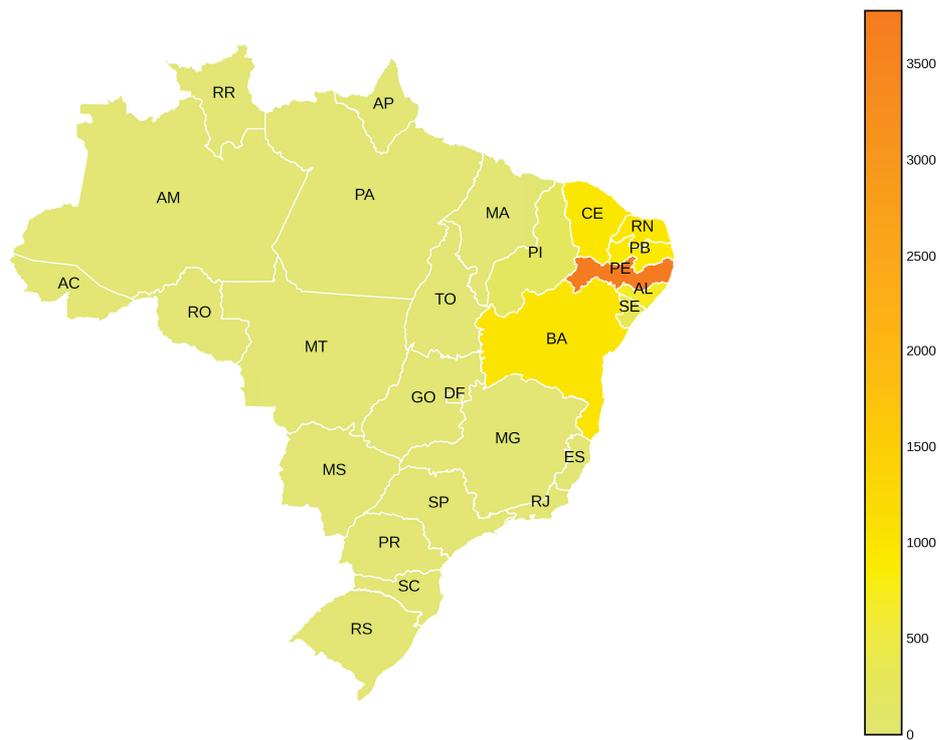
Figura 40 – Análise da quantidade de menções aos termos das categorias Consumidores reais e Potenciais consumidores referentes ao ato de consumir queijos artesanais em conjunto com seus acompanhamentos e receitas. **Fonte:** autor.

#### 4.3.2.1 Análise do Perfil de Consumo do Queijo Coalho

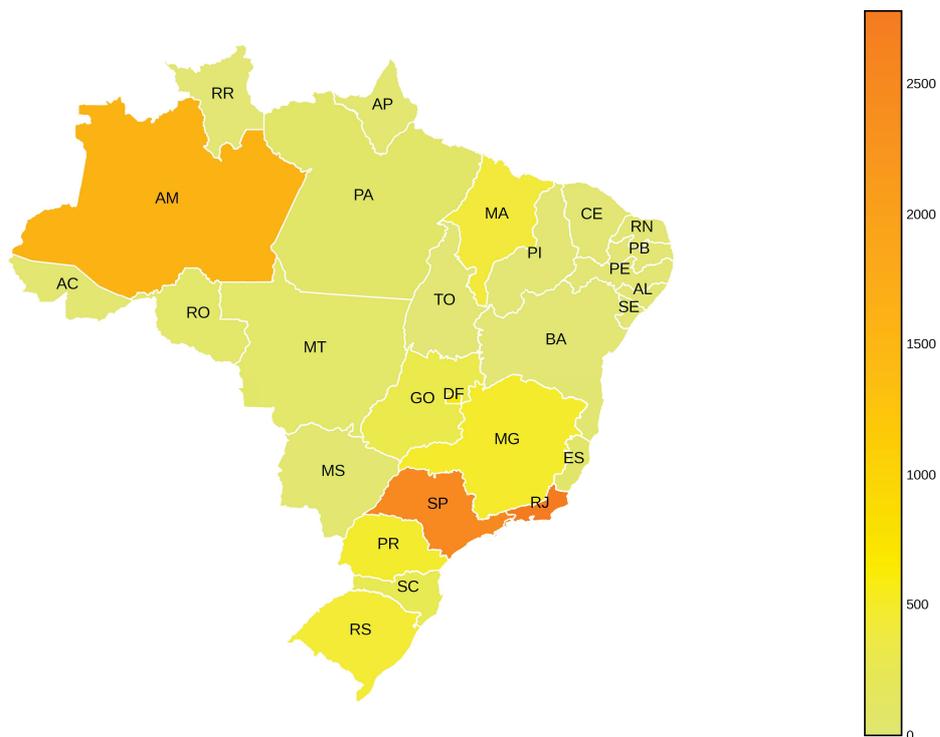
Agora, para mostrar o potencial da ferramenta do Observatório do Consumidor, dado que já foi mostrado um panorama geral dos queijos artesanais, o intuito é o de realizar uma análise mais focada em um só tipo de queijo, que neste caso foi o **Queijo Coalho**, por ter sido o queijo com maior número de *tweets*. Foram publicados 43.339 *tweets* sobre queijo coalho no Brasil. Desse total, 19.071 *tweets* possuíam a localização ativada, o que permitiu realizar uma análise separada do queijo coalho tradicional ou regional do Nordeste. Assim, foram estudados 8.873 *tweets* provenientes da região Nordeste, representando 46,5% do total de postagens sobre queijo coalho no Brasil. Os restantes (10.198 postagens ou 53,5% dos *tweets*) eram provenientes das demais regiões. Desta maneira, a Figura 41 mostra estes resultados por regiões.

No Nordeste, Pernambuco foi o estado que apresentou o maior número de publicações sobre queijo coalho e também a maior variação de preços para o produto. O maior número de *posts* referia-se a acompanhamentos e receitas que empregam o queijo coalho (1.451). Dentre essas receitas, as mais citadas foram: tapioca (27,3%), carne de sol (23,5%), pizza (9,7%) e arroz (7,3%). Nas postagens que envolviam a tapioca, geralmente havia a combinação desta com queijo coalho e outros produtos que remetem à indulgência, como por exemplo, mel e doce de leite. Já em relação aos acompanhamentos para este tipo de queijo, a carne (14,1%) e o pão (9,8%) foram os mais citados enquanto o café (38%), seguido de leite (28,2%), suco (13,2%), cerveja (6,5%) e chá (6%) representaram as bebidas que frequentemente acompanharam o queijo tipo coalho. Com relação à textura, a maioria das postagens fazia referência à cremosidade (35%) e com relação à composição, o orégano foi o ingrediente mais citado, referindo-se principalmente ao consumo de dadinhos de queijo coalho com orégano.

O conteúdo publicado no Twitter sobre queijo coalho nas demais regiões do Brasil apresentou algumas similaridades com o conteúdo publicado no Nordeste. Nas demais regiões, o maior número de postagens também foi de acompanhamentos e receitas. Para os acompanhamentos, o pão ficou em primeiro lugar com 22,3% das citações, seguido por carne (12,9%) e tucumã (7,4%). Para as receitas, novamente a tapioca ficou em primeiro colocado, com 29,9% das postagens, seguido por carne sol com 11,7%, pizza com 8,7% e arroz com 7,7% das citações. Porém, as associações da tapioca com o queijo coalho nas demais regiões do País foram acompanhadas de manteiga ou banana e tucumã, o que difere do consumo indulgente observado para o queijo coalho no Nordeste. No quesito bebidas para acompanhar o queijo coalho nas 4 regiões do Brasil (Norte, Sul, Sudeste e Centro-Oeste), o café foi novamente o mais citado com 28,6% das menções. Na sequência, observou-se leite (21,1%), cerveja (15,3%) e suco (14,1%). No entanto, não foi possível analisar outros atributos relacionados ao consumo deste tipo de queijo devido ao reduzido número de publicações a este respeito.



(a) Região nordeste



(b) Demais Estados

Figura 41 – Regiões do Brasil que mencionaram o Queijo Coalho. Em (a) é exibido somente os resultados referentes ao número de publicações sobre Queijo Coalho na região nordeste atribuindo 0 a todos os demais estados. Em (b) é exibido somente os resultados referentes ao número de publicações sobre Queijo Coalho nas demais regiões atribuindo 0 a todos os estados referentes à região nordeste do Brasil. **Fonte:** autor.

Portanto, nota-se uma tendência de consumo indulgente do queijo coalho na região Nordeste. Apesar do estudo não ter avaliado os momentos das refeições, pode-se inferir que o queijo coalho é consumido preferencialmente no café da manhã ou no lanche da tarde, visto que pão e café apareceram como acompanhamentos muito citados. Além disso, tem o leite também que é a bebida mais relacionada a este tipo de refeição. As demais citações, como arroz, carne de sol, pizza, etc, podem se referir tanto a almoço quanto a jantar. A tapioca também pode ser consumida em qualquer refeição, inclusive como sobremesa. Isso indica que o queijo coalho apresenta grande versatilidade de consumo no Brasil. Apesar das diferenças regionais, o produto tem sido consumido amplamente nas mais diversas formas, refeições e acompanhamentos. Com isso, é possível entender a sua popularidade de Norte a Sul do País.

#### 4.4 TELAS DA PLATAFORMA DO OBSERVATÓRIO DO CONSUMIDOR

Esta seção tem como objetivo exibir as telas da plataforma do Observatório do Consumidor desenvolvidas referentes à primeira versão da ferramenta. A construção desta ferramenta foi realizada por estagiários do Programa de Residência Zootécnica Digital da EMBRAPA Gado de Leite em parceria com a Universidade Federal de Juiz de Fora e o Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas - Campus Juiz de Fora (IF-Sudeste/MG).



# Queijos Artesanais

Para ser considerado artesanal, o queijo deverá atender aos requisitos presentes no Decreto nº 9.510/2019. Abaixo encontram-se relacionados os principais tipos de queijos artesanais consumidos no mercado.

## Consulta

As consultas são realizadas por meio da aplicação de mineração de dados. A mineração usa técnicas estatísticas e de Aprendizado de Máquina para analisar o comportamento dos usuários em relação aos queijos artesanais.

[Veja mais ▶](#)

## Projeto

O Observatório do Consumidor atualmente fornece ao usuário uma interface que o possibilita visualizar de forma intuitiva identificar opiniões, características, hábitos e tendências referente ao consumo de queijos artesanais no Brasil.

[Veja mais »](#)



Figura 42 – Tela inicial da plataforma do Observatório do Consumidor. **Fonte:** autor.

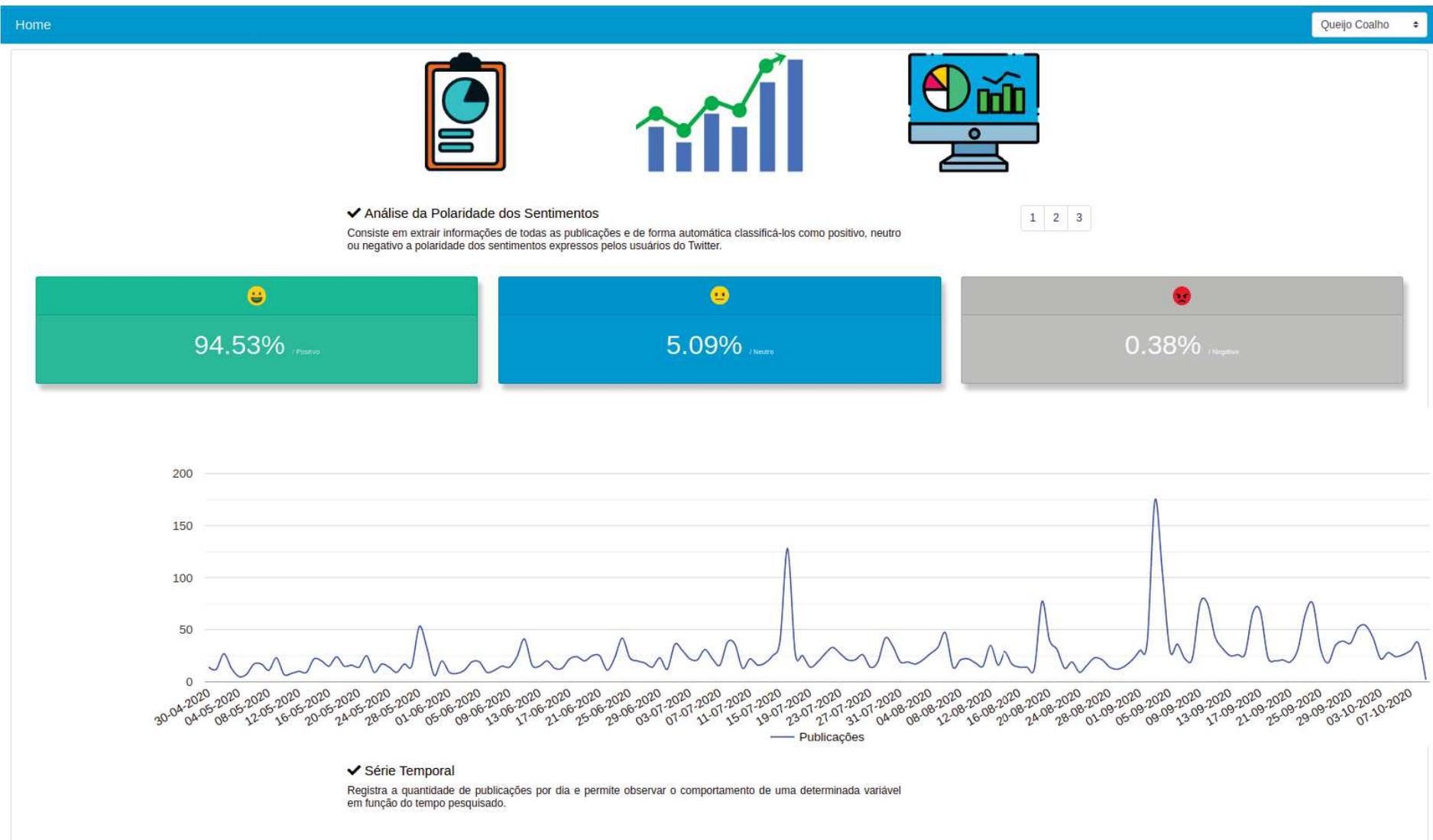
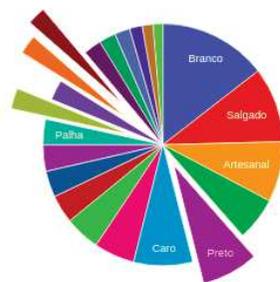


Figura 43 – Tela com os resultados da análise de sentimentos e análise temporal da quantidade de *tweets* ao longo do tempo. **Fonte:** autor.

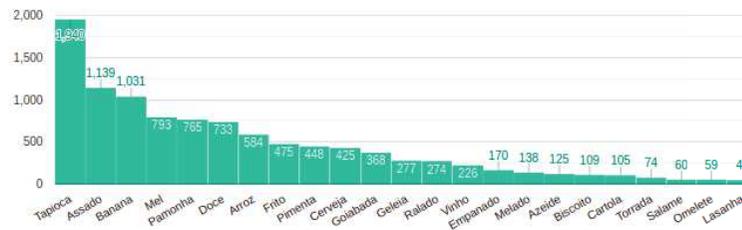
### Características do Queijo Artesanal



#### ✓ Característica do Queijo

O gráfico apresenta dados referentes às principais características relatadas por usuários da rede, vinculadas a categoria pesquisada.

### Hábitos de Consumo



#### ✓ Hábitos de consumo do queijo artesanal no Brasil

O gráfico de hábitos de consumo relaciona os principais alimentos que são consumidos com a categoria pesquisada.

Figura 44 – Tela com os resultados das características e hábitos de consumo do queijo artesanal. **Fonte:** autor.

## 5 CONCLUSÃO

Seguindo o que propõe o *Business Intelligence* (BI), neste trabalho, foi possível desenvolver um sistema computacional denominado “Observatório do Consumidor” que se apresenta como uma ferramenta capaz de coletar, organizar, processar e transformar dados do Twitter em informação, auxiliando na tomada de decisões em ambiente corporativo, identificando oportunidades e aumentando as chances de sucesso de uma determinada ação, principalmente com relação ao tema abordado neste trabalho que são os queijos artesanais.

Foi possível desenvolver um banco de dados relacional específico para o sistema computacional do “Observatório do Consumidor”, que consegue armazenar de forma eficiente tanto os dados coletados do Twitter quanto os resultados gerados pelos processamentos realizados pela ferramenta.

Foram realizadas análises exploratórias nos dados coletados, em que foi possível analisar a quantidade de *tweets* publicados sendo 82.959 *tweets* no total com uma média de 1.997 *tweets* por semana. Também foi possível verificar quais foram as categorias de tipos de queijos artesanais mais citados no Twitter, sendo a do Queijo Coalho a de maior destaque com 52,2% dos *tweets*. Outra análise possível foi referente os estados brasileiros que mais publicaram sobre os queijos artesanais. Cerca de 22.635 *tweets* continham esta informação sendo os estados de São Paulo, Rio de Janeiro, Minas Gerais, Pernambuco, Rio Grande do Sul, Amazonas, Distrito Federal, Bahia, Rio Grande do Norte e Paraná, os 10 estados que mais publicaram representando 79,4% do total de localizações válidas.

Utilizando técnicas de processamento de linguagem natural (PLN) foi possível desenvolver algoritmos para tratamento dos dados textuais do Twitter. Realizando o processo de tokenização, foi possível segmentar o texto dos *tweets* dos usuários através do delimitador espaço em branco e remover as *stopwords* retirando os termos que não trazem informações relevantes. Em seguida para reduzir a redundância de palavras parecidas nos dados utilizou-se o processamento de *stemming* que reduz cada palavra ao seu radical comum. Após, foi utilizado a vetorização por TF-IDF para a criação de um vetor de frequência dos termos presentes nos textos finalizando com a técnica de n-gramas, que cria um termo com duas ou mais palavras a fim de facilitar a análise de palavras compostas.

Utilizando uma abordagem supervisionada, foi possível desenvolver um modelo de análise da polaridade dos sentimentos dos *tweets* utilizando aprendizagem de máquina supervisionada, a fim de identificar a polaridade dos sentimentos presentes nos conteúdos dos *tweets* publicados pelos usuários do Twitter, verificando se o sentimento presente era negativo, neutro ou positivo ao mencionarem os queijos artesanais em suas publicações. Com respeito a polaridade do sentimento presente nos *tweets*, a maioria possui um sentimento positivo mostrando a prevalência de um público que emite uma opinião ou

sensação positiva com relação ao queijo artesanal na rede social do Twitter. Tal informação é de extrema utilidade para pequenos, médios e grandes produtores de queijo artesanal no Brasil, pois, de um modo geral, foi possível verificar que grande parte dos consumidores que possuem uma conta no Twitter, fizeram publicações com um sentimento positivo.

Após a construção das categorias que representam as características e hábitos de consumo do queijo artesanal, foi possível implementar técnicas de mineração de dados a fim de realizar análises de conteúdo nos *tweets*. Verificou-se que definir uma tendência geral para todos os queijos artesanais no Brasil foi impraticável devido a questões como o regionalismo, características e hábitos de consumo que variam para cada tipo de queijo, fato este que pôde ser observado pela análise do perfil de consumo do queijo Coalho.

Contudo, é possível concluir que através da implementação da ferramenta do “Observatório do Consumidor” diversas análises foram realizadas possibilitando identificar informações interessantes sobre o cenário das tendências de consumo dos queijos artesanais no Brasil. É relevante destacar que esta ferramenta não se limita exclusivamente aos queijos artesanais, podendo ser estendida a outros produtos. As principais contribuições que esta solução permite são as de combinar diversas técnicas e metodologias em um só lugar, tornando a tomada de decisão mais eficiente pautada pela análise de dados reais, inovando a forma de fazer pesquisas de mercado tradicionais 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 - do queijo artesanal no Brasil.

## 6 TRABALHOS FUTUROS

Como trabalhos futuros, pretende-se ampliar o universo de produtos a serem analisados, em que espera-se englobar diversas categorias de derivados lácteos. Pretende-se também promover melhorias na base de treinamento do modelo de análise da polaridade de sentimento, e realizar adaptações no banco de dados e construção de novos algoritmos de coleta e análise de dados, a fim de potencializar os resultados alcançados até o presente momento, bem como outros que ainda virão após a implementação de novas análises.

Outra perspectiva é a de expandir a análise de sentimentos para a análise de emoções com o objetivo de identificar alegria, tristeza, raiva, nojo, medo e surpresa, presente nas publicações das redes sociais. Para isso será necessário melhorar e adaptar a base de treinamento do modelo, mantendo um balanceamento entre as classes a fim de obtermos melhores resultados de acurácia, do coeficiente de determinação  $R^2$ , do MSE, do RMSE e MAPE.

Outra melhoria futura, é a ampliação da coleta de dados para outras redes sociais, em princípio Facebook e Instagram, com o intuito de obter outras fontes e tipos de dados, tornando nossa ferramenta mais completa e abrangente. Adicionalmente, com os dados do Instagram será proposto a implementação de algoritmos de visão computacional para identificação de características e hábitos de consumo dos derivados lácteos em geral presentes nas imagens publicadas pelos usuários/consumidores.

## REFERÊNCIAS

- ABEL, Carol. **Exemplos de Business Intelligence: 5 casos para se inspirar**. 2018. Disponível em: <https://bityli.com/4Fngt>.
- AI, GOTIT. **GOTIT - Sentiment**. 2021. Disponível em: <https://bityli.com/79MH7>.
- ANDRANIK, T.; SPRENGER, T.; P., Sander; WELPE, I. Predicting elections with twitter: What 140 characters reveal about political sentiment. in Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010), p. 178–185, 2010.
- ANGELONI, Maria T.; REIS, Eduardo S. Business intelligence como tecnologia de suporte a definição de estratégias para melhoria da qualidade do ensino. **Encontro da ANPAD.**, p. 16, 2006.
- ANKIT; SALEENA, Nabizath. An ensemble classification system for twitter sentiment analysis. **Procedia Computer Science**, v. 132, p. 937–946, 2018. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science. Disponível em: <https://www.sciencedirect.com/science/article/pii/S187705091830841X>.
- APROALAGOA. **Um Queijo e muitas Histórias**. 2021. Disponível em: <https://bityli.com/DihcC>.
- AVELAR, Cátia; ROCHA, Thiago; CRUZ, Flávia. MineraÇÃO de dados. **Revista Vianna Sapiens**, v. 8, p. 25, 12 2017.
- AYDANO p. Machado. Mineração de texto em redes sociais aplicada à educação a distância. 2010. Disponível em: <http://pead.ucpel.tche.br/revistas/index.php/colabora/article/view/132>.
- BARABBA, T.; ZALTAMAN, P. Hearing the voice of the market. **Harvard Business School Press**, 1991.
- BARBOSA, L.; L., Madi; A., Toledo M.; REGO, R. A. Tendências da alimentação. **Brasil Food Trends**, 2020.
- BARR, A.; FEIGENBAUM, E. A. The handbook of artificial intelligence. **Addison-Wesley Publishing Company**, I, 1981.
- BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. **Short course in the Brazilian Symposium on Multimedia and the Web (Webmedia)**, 2015.
- BERNARDO, ANDRÉ. **BBC Brasil: Os bastidores da viagem de 44 dias que levou Pedro Álvares Cabral ao Brasil**. 2020. Disponível em: <https://bityli.com/9KcxV>.
- BOARI, Cleube Andrade. **História dos queijos artesanais no Brasil**. 2017. Disponível em: <https://bityli.com/frSEv>.
- BRASIL. **REGULAMENTO TÉCNICO DE IDENTIDADE E QUALIDADE DE QUEIJO DE MANTEIGA**. 2001. Disponível em: <https://bityli.com/cHw3W>.
- BRASIL. Lei n13.860 de 18 de julho de 2019. 2019.

- CASTRO, R. **Queijo poronguinho**. 2013. Disponível em: <https://bitly.com/eRk7X>.
- CAVALCANTE, J. F. M. Sistema de apoio à decisão na produção de leite e queijo coalho com segurança alimentar. **Tese (Doutorado em Ciência e Tecnologia dos Alimentos)**. Universidade Federal de Viçosa., 2005.
- CAVALCANTE, J. F. M.; et al. Processamento do queijo coalho regional empregando leite pasteurizado e cultura láctica endógena. *Ciênc. Tecnol. Aliment.*, 2007.
- CAÇADOR, Câmara Municipal de. **Projeto autoriza produtores locais a comercializarem queijo colonial**. 2017. Disponível em: <https://bitly.com/p8dui>.
- CHAVES, G.; DOLORES, F.; AGUIAR, V. **Sertão Nordestino: A região sertaneja e agreste do Nordeste tem na carne-seca e de bode os principais ingredientes de sua rica cozinha**. 2021. Disponível em: <https://bitly.com/xMTDG>.
- COMERQUEIJO. **Prêmio Queijo Brasil**. 2019. Disponível em: <https://bitly.com/h84kd>.
- COPELAND, Michael. **What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?** 2016. Disponível em: <https://bitly.com/esfi2>.
- COUTINHO, L.; CANEPPELLE, V.D.C.; DIAS, K.; DEMOZZI, C.; CONTO, L.C. Análise descritiva quantitativa (adq) de queijo do tipo serrano. 2018.
- COVINGTON, M. Nlp for prolog programmers. **Prentice-Hall**, 1994.
- DELHAYE, Heaven. **Queijo Colonial**. 2016. Disponível em: <https://bitly.com/jy45s>.
- DEVELOPER, Inc. 2021 TWITTER. **Twitter Object**. 2021. Disponível em: <https://bitly.com/NHjNu>.
- DIAS, M.M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. **ActaScientiarum**, p. 1715–1725, 2002.
- DORES, M. T. das. Enterotoxigenicidade de staphylococcus aureus isolados de queijo minas artesanal da canastra. p. 66, 2013.
- DUDA, R. O.; HART, P. E.; STORK, D. G. Unsupervised learning and clustering.: Pattern classification. **Chichester**, p. 519–598, 2001.
- FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. de. Inteligência artificial: uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, 2011.
- FINZI, Eduardo. **Linguagem natural: entenda o que é e a importância na era de chatbots**. 2019. Disponível em: <https://bitly.com/I6JkA>.
- FOOD, Brasil Slow. **Queijo Artesanal de Leite Cru**. 2018. Disponível em: <https://bitly.com/pVNMs>.
- FOUNDATION, Python Software. **Python Language Reference. Version 3.7**. 2021. Disponível em: <https://bitly.com/zMnhv>.

- FRANCO, M. **Requeijão de corte: uma iguaria bem mineira**. 2018. Disponível em: <https://bityli.com/BzxIS>.
- GALVÃO, N.D.; MARIN, H.F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, p. 686–90, 2009.
- GARTNER. **Definition of Big Data**. 2021. Disponível em: <https://bityli.com/ZYfJn>.
- GONZALEZ, Leandro de Azevedo. **Regressão logística e suas aplicações**. 2018.
- GORETTI, Priscila. **Tapioca emagrece? Saiba como ela pode ser usada para ajudar na perda de peso**. 2021. Disponível em: <https://bityli.com/IMGNE>.
- GROUP, Kelsey. **Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior, Machine Learning, and Deep Learning?** 2007. Disponível em: <https://bityli.com/qgzFK>.
- HAIR, JR. J. F.; BUSH, R. P.; ORTINAU, D. J. Marketing research within a changing information environment. **Boston: McGraw-Hill/Irwin series in marketing**, 2003.
- HAT, Red. **O que é open source?** 2021. Disponível em: <https://bityli.com/QJ1zs>.
- HIPPISLEY, A. **Lexical analysis**: In handbook of natural language processing. 2nd edition. ed. [S.l.]: Chapman and Hall/CRC, 2010.
- IBGE. **Biomass**. 2021. Disponível em: <https://bityli.com/TILoe>.
- IPHAN. **Modo Artesanal de Fazer o Queijo Minas: Serro, Serra da Canastra e Serra do Salitre (Alto Paranaíba)/Instituto do Patrimônio Histórico e Artístico Nacional**. 2014. Disponível em: <https://bityli.com/Bdaxr>.
- ITTNER, A. **Queijo Kochkäse: portaria regulamenta produção e venda da iguaria em Santa Catarina**. 2020. Disponível em: <https://bityli.com/gTIKQ>.
- JESÚS, P. M. **Modelando a linguagem com N-gramas**. 2019. Disponível em: <https://bityli.com/Iv5Qe>.
- KEMP, Simon. **Digital 2019**. 2019. Disponível em: <https://bityli.com/5ePzB>.
- KEMP, Simon. **Digital 2020: Brazil**. 2020. Disponível em: <https://bityli.com/GaA6j>.
- LAVOURA, A. **Produzido com leite de búfala, Queijo do Marajó é único no mundo**. 2020. Disponível em: <https://bityli.com/0YTky>.
- LIU, B. Sentiment analysis and opinion mining.: Synthesis lectures on human language technologies. **Morgan Claypool Publishers.**, v. 5, p. 1–167, 2012.
- LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: **In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics**. Philadelphia: Association for Computational Linguistics. [S.l.: s.n.], 2002.
- LUGER, G. F. Inteligência artificial: estruturas e estratégias para a solução de problemas complexos. 4° ed. **Bookman**, 2004.

MAIMON, O.; ROKACH, L. Data mining and knowledge discovery handbook. **Springer**, v. 2, 2005.

MANDACARU. **Queijo de Coalho Mandacaru**. 2021. Disponível em: <https://bitly.com/ar6UB>.

MANDARINO, D. **Conheça as diferenças entre os queijos colonial e serrano**. 2018. Disponível em: <https://bitly.com/JTh2s>.

MARESCH, G. **Queijinho Branco e Kochkäse: conheça os Queijos Artesanais e Coloniais produzidos em Santa Catarina**. 2020. Disponível em: <https://bitly.com/12gxi>.

MCGLOHON, M.; GLANCE, N.; REITER, Z. Star quality: Aggregating reviews to rank products and merchants. in Proceedings of the International Conference on Weblogs and Social Media (ICWSM- 2010), p. 114–121, 2010.

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.

MELO, R.; RODRIGUES, F.; SOARES, L. **Queijos de MG conquistam 50 medalhas em concurso mundial na França**. 2019. Disponível em: <https://bitly.com/QDYEG>.

MILKPOINT, Giro de Notícias. **PR: produtores de queijo de cabra aumentam oferta para atender consumo em alta**. 2020. Disponível em: <https://bitly.com/0A0BT>.

MJV, Team. **Redes Sociais e Big Data: a melhor estratégia para conhecer seus clientes**. 2020. Disponível em: <https://bitly.com/hUdNr>.

MOHAMED, Sinkadar. **100 Social Media Statistics You Must Know [2018] + Infographic**. 2018. Disponível em: <https://bitly.com/IHY4V>.

MONTEIRO, R. P. Capítulo 1 - o queijo minas artesanal e seu potencial para a agroindústria familiar. 2018.

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

MYSQLLAB. **MySQL Database**. 1995. Disponível em: <https://bitly.com/lNGqY>.

NARKHEDE, Sarang. **Understanding Confusion Matrix**. 2018. Disponível em: <https://bitly.com/tSNzN>.

OFICINA, Redação. **Por que usar MySQL?** 2007. Disponível em: <https://bitly.com/hllGa>.

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

PAULISTA, ARTESANAL. **Caminho do Queijo Artesanal Paulista**. 2021. Disponível em: <https://bitly.com/KjpsY>.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O. Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, p. 2825–2830., 2011.

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

PIRES, E. F.; MORAIS, C. M. M.; SILVA, J. A.; CARVALHO, R. O. C. Queijo de coalho – perfil industrial. **SEBRAE**, v. 1, p. 44, 1994.

QUEIJOS, Sobre. **Queijo Cabacinha: curiosidades e fabricação**. 2020. Disponível em: <https://bitly.com/6Bo1S>.

RAISINGHANI, M. Business intelligence in the digital economy. **Hershey PA:The Idea Group**, 2004.

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

RODOLFHO, S. **IMA - Assessoria de Comunicação: Regiões produtoras de queijos artesanais são reconhecidas pelo Governo de Minas**. 2020. Disponível em: <https://bitly.com/6MlmB>.

ROSENFELD, G. H.; FITZPATRICK, K. A coefficientt of agreement as a measure of thematic classification accuracy. **Photogrammetric Engineering and Remote Sensing.**, v. 52, n. 2, p. 227–233, 1986.

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

SANTANA, Marlesson. **Deep Learning: do Conceito às Aplicações**. 2018. Disponível em: <https://bitly.com/r8ORP>.

SHARMA, Aditya. **Differences Between Machine Learning Deep Learning**. 2018. Disponível em: <https://bitly.com/OJVS5>.

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

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

SIQUEIRA, K. B.; ROCHA, D. T. **Consumo de lácteos na pandemia: uma análise das variações de consumo entre as classes de renda e regiões no Brasil**. CILeite, 2020. Disponível em: <https://bitly.com/28zYg>.

SOARES, N.; GOLIATT, L.; SANTIS, R.; E.P., Aguiar. Máquinas de aprendizado extremo para a predição do volume de clones de eucaliptos. **Conference: Ibero-Latin American Congress on Computational Methods in Engineering - CILAMCE At: Florianópolis, SC, 2017.**

SOUZA, A. **Requeijão do Sertão, uma tradição ameaçada.** 2016. Disponível em: <https://bityli.com/aPvuG>.

SOUZA, Lucas Alves Moreira de. Aplicação de aprendizado de máquina para predição de prioridade em gestão de incidentes. 2017.

TABLEAU. **O que é business intelligence? Seu guia sobre o BI e por que ele é importante.** 2021. Disponível em: <https://bityli.com/WV4Pg>.

TERRA, Carolina Frazon. Mídia social: fator de influência para o consumo? um estudo com usuários de redes sociais sobre os critérios de influência na hora da compra. **Comunicom**, 2012. Disponível em: <https://bityli.com/uwoEN>.

TORRES, Thaila. **Projeto quer resgatar história e tradição do queijo nicola, de origem pantaneira.** Campo Grande News, 2017. Disponível em: <https://bityli.com/tVt0h>.

TRENDS, Google. **Google Trends.** 2021. Disponível em: <https://bityli.com/wsTn1>.

TURBAN, E.; SHARDA, R.; DELEN, D. **Decision Support and Business Intelligence Systems.** [S.l.: s.n.], 2010. p.720 p.

TWITTER, Inc. 2021. **Twitter.** 2021. Disponível em: <https://bityli.com/5xGla>.

VERÍSSIMO, Bruna; LEPRE, Larissa; TINCANI, Daniela. Diferenças entre pesquisa de marketing e pesquisa de neuromarketing. 2018.

WAGNER, S. A. Desenvolvimento do rtiq do queijo serrano: resultados e desafios. 2018.

WASSERMAN, S.; FAUST, k. Social network analysis: methods and applications. **Cambridge University Press.**, 1994.

WILLIAM, Scott. **TF-IDF from scratch in python on real world dataset.** 2019. Disponível em: <https://bityli.com/F2key>.

WITTEN, Ian H.; FRANK, Eibe. Data mining practical machine learning tools and techniques. Elsevier, 2005.

WORLDOMETERS. **World Population Clock: 7.6 Billion People.** 2017. Disponível em: <https://bityli.com/YxWg0>.

ZACARCHENCO, P. B.; DENDER, A. G. F. V.; REGO, R. A. Brasil dairy trends 2020: Tendências de mercado de produtos lácteos. 2020.

## APÊNDICE A – LISTA DE PALAVRAS-CHAVE

Tabela 6 – Lista de palavras-chave usadas para coleta no Twitter.

<b>Id</b>	<b>Palavra-chave</b>	<b>Categoria do Queijo</b>
1	parmesao da mantiqueira	Alagoa
2	queijo alagoa	Alagoa
3	queijo alagoa fumace	Alagoa
4	queijo alagoa pequena	Alagoa
5	queijo alagoinha	Alagoa
6	queijo artesanal fazenda jm	Alagoa
7	queijo cirico	Alagoa
8	queijo do coronel	Alagoa
9	queijo do inacio	Alagoa
10	queijo do jayr	Alagoa
11	queijo faixa dourada	Alagoa
12	queijo famoso de alagoa	Alagoa
13	queijo fazenda bela vista	Alagoa
14	queijo fazenda corguinho	Alagoa
15	queijo fazenda santo antonio	Alagoa
16	queijo garrafao	Alagoa
17	queijo garrafao sitio mr	Alagoa
18	queijo gir	Alagoa
19	queijo padre victor	Alagoa
20	queijo paschoal poppa	Alagoa
21	queijo pe da pedra	Alagoa
22	queijo pousada fazenda corguinho	Alagoa
23	queijo sabor da alagoa	Alagoa
24	queijo santa inacia	Alagoa
25	queijo serra do condado	Alagoa
26	queijo sitio pe da pedra	Alagoa
27	queijo azul	Artesanal Paulista
28	queijo azul do bosque	Artesanal Paulista
29	queijo braukase	Artesanal Paulista
30	queijo capao	Artesanal Paulista
31	queijo carijo	Artesanal Paulista
32	queijo creme	Artesanal Paulista
33	queijo cuesta	Artesanal Paulista

*Continua na próxima página*

Tabela 6 – *Continua na próxima página*

<b>Id</b>	<b>Palavra-chave</b>	<b>Categoria do Queijo</b>
34	queijo cuesta azul	Artesanal Paulista
35	queijo cuestinha	Artesanal Paulista
36	queijo dona carolina	Artesanal Paulista
37	queijo garnize	Artesanal Paulista
38	queijo giramundo	Artesanal Paulista
39	queijo granito	Artesanal Paulista
40	queijo lua	Artesanal Paulista
41	queijo mandala	Artesanal Paulista
42	queijo mantiqueira da serra	Artesanal Paulista
43	queijo mimo da serra	Artesanal Paulista
44	queijo mogiana	Artesanal Paulista
45	queijo pardinho artesanal	Artesanal Paulista
46	queijo paulista	Artesanal Paulista
47	queijo piramide do bosque	Artesanal Paulista
48	queijo pora	Artesanal Paulista
49	queijo provola de caverna	Artesanal Paulista
50	queijo simmental	Artesanal Paulista
51	queijo soberano	Artesanal Paulista
52	queijo sol do japi	Artesanal Paulista
53	queijo tropeirinho	Artesanal Paulista
54	queijo tropeiro	Artesanal Paulista
55	queijo tulha	Artesanal Paulista
56	queijo cabacinha	Cabacinha
57	queijo coalho	Coalho
58	queijo coalho de sol	Coalho
59	queijo coalho santana	Coalho
60	queijo coalho natural	Coalho
61	queijo coalho de bufala	Coalho
62	queijo colonial	Colonial
63	queijo diamante	Colonial
64	queijo contestado	Contestado
65	queijo kochkase	Kochkäse
66	queijo manteiga	Manteiga
67	queijo manteiga com raspas	Manteiga
68	queijo cataua	Mantiqueira de Minas

*Continua na próxima página*

Tabela 6 – *Continua na próxima página*

<b>Id</b>	<b>Palavra-chave</b>	<b>Categoria do Queijo</b>
69	queijo goa	Mantiqueira de Minas
70	queijo do marajo	Marajó
71	queijo bello	Maturado
72	queijo do cerrado	Maturado
73	queijo liberdade	Maturado
74	queijo cabra	Maturado
75	queijo maturado	Maturado
76	queijo azul de ovelha	Maturado
77	queijo araxa	Minas Artesanal
78	queijo barreiro	Minas Artesanal
79	queijo bicas da serra	Minas Artesanal
80	queijo bom sucesso	Minas Artesanal
81	queijo canaa	Minas Artesanal
82	queijo canastra	Minas Artesanal
83	queijo canastra casquinha	Minas Artesanal
84	queijo canastra onesio	Minas Artesanal
85	queijo canastra raiz	Minas Artesanal
86	queijo canastra real	Minas Artesanal
87	queijo canastra reinaldo	Minas Artesanal
88	queijo canastra serjao	Minas Artesanal
89	queijo canastra solange	Minas Artesanal
90	queijo capao grande	Minas Artesanal
91	queijo capim canastra	Minas Artesanal
92	queijo carranca	Minas Artesanal
93	queijo curupira	Minas Artesanal
94	queijo da marisa	Minas Artesanal
95	queijo dama de araxa	Minas Artesanal
96	queijo do cedro	Minas Artesanal
97	queijo do dinho	Minas Artesanal
98	queijo do gilson	Minas Artesanal
99	queijo do ivair	Minas Artesanal
100	queijo do johne	Minas Artesanal
101	queijo do juquinha	Minas Artesanal
102	queijo do miguel	Minas Artesanal
103	queijo do rei	Minas Artesanal

*Continua na próxima página*

Tabela 6 – *Continua na próxima página*

<b>Id</b>	<b>Palavra-chave</b>	<b>Categoria do Queijo</b>
104	queijo do reinaldo	Minas Artesanal
105	queijo do serjao	Minas Artesanal
106	queijo do serjao merendeiro	Minas Artesanal
107	queijo do tino	Minas Artesanal
108	queijo do tulio	Minas Artesanal
109	queijo dona iaia	Minas Artesanal
110	queijo fazenda beira da serra	Minas Artesanal
111	queijo fazenda capao grande	Minas Artesanal
112	queijo ferreira campos	Minas Artesanal
113	queijo gurita	Minas Artesanal
114	queijo guzera	Minas Artesanal
115	queijo ivituruy	Minas Artesanal
116	queijo jua	Minas Artesanal
117	queijo luana	Minas Artesanal
118	queijo merendeiro	Minas Artesanal
119	queijo militao da canastra	Minas Artesanal
120	queijo minas araxa	Minas Artesanal
121	queijo minas artesanal	Minas Artesanal
122	queijo minas artesanal serro	Minas Artesanal
123	queijo minas canastra	Minas Artesanal
124	queijo minas estrada real	Minas Artesanal
125	queijo minas salitre	Minas Artesanal
126	queijo minas serro	Minas Artesanal
127	queijo moldura de minas	Minas Artesanal
128	queijo onca	Minas Artesanal
129	queijo ouro fino	Minas Artesanal
130	queijo paiol velho	Minas Artesanal
131	queijo pingo de amor	Minas Artesanal
132	queijo pingo de mula	Minas Artesanal
133	queijo porto canastra	Minas Artesanal
134	queijo quilombo	Minas Artesanal
135	queijo rancho 4r	Minas Artesanal
136	queijo recanto dos pimentas	Minas Artesanal
137	queijo santana	Minas Artesanal
138	queijo santuario do mergulhao	Minas Artesanal

*Continua na próxima página*

Tabela 6 – *Continua na próxima página*

<b>Id</b>	<b>Palavra-chave</b>	<b>Categoria do Queijo</b>
139	queijo senzala	Minas Artesanal
140	queijo serra dos araxas	Minas Artesanal
141	queijo serro	Minas Artesanal
142	queijo sitio talisma	Minas Artesanal
143	queijo taroco	Minas Artesanal
144	queijo terrua das vertentes	Minas Artesanal
145	queijo tradicao da canastra	Minas Artesanal
146	queijo tres ranchos	Minas Artesanal
147	queijo trilha da pedra	Minas Artesanal
148	queijo turvo grande	Minas Artesanal
149	queijo vo cotinha	Minas Artesanal
150	quejo tradicao da canastra	Minas Artesanal
151	queijo minas campos das vertentes	Minas Artesanal
152	queijo mini terrua das vertentes	Minas Artesanal
153	queijo nicola	Nicola
154	queijo capril rancho das vertentes	Outros
155	queijo raclette mineiro	Outros
156	queijo 0 lactose	Outros
157	queijo artesanal	Outros
158	queijo bufala	Outros
159	queijo curado	Outros
160	queijo fungo	Outros
161	queijo gourmet	Outros
162	queijo meia cura	Outros
163	queijo minas cabra	Outros
164	queijo mofo	Outros
165	queijo mofo branco	Outros
166	queijo ovelha	Outros
167	queijo parmesao artesanal	Outros
168	queijo roca	Outros
169	queijo sem lactose	Outros
170	queijo tradicional	Outros
171	queijo zero lactose	Outros
172	queijos artesanais	Outros
173	queijo grupiara	Outros

*Continua na próxima página*

Tabela 6 – *Continua na próxima página*

<b>Id</b>	<b>Palavra-chave</b>	<b>Categoria do Queijo</b>
174	queijo porongo	Porongo
175	queijinho	Queijinho
176	requeijao moreno	Requeijão
177	requeijao do sertao	Requeijão
178	queijo artesanal serrano	Serrano
179	queijo serrano	Serrano

**Fonte:** autor.

## APÊNDICE B – DEFINIÇÕES DOS ATRIBUTOS

Segundo o site de desenvolvedor do Twitter (DEVELOPER, 2021), ao coletar os *tweets* é retornado um objeto chamado *tweet* que possui diversas informações, sendo as descritas abaixo utilizadas neste trabalho.

- **user\_id\_str**: informação referente ao identificador único do usuário na plataforma do Twitter;
- **user\_name**: informação referente ao nome do usuário na plataforma do Twitter;
- **user\_screen\_name**: apelido dado pelo usuário na plataforma do Twitter;
- **user\_location**: exibe a localização que o usuário disponibilizou na plataforma do Twitter;
- **user\_url**: se refere ao endereço url do usuário na plataforma do Twitter;
- **user\_description**: descrição pessoal disponibilizada pelo usuário na plataforma do Twitter;
- **user\_followers\_count**: contador de seguidores que um determinado usuário possui;
- **user\_friends\_count**: referente à quantidade de contas de usuários que uma determinada conta está seguindo;
- **user\_listed\_count**: quantidade de listas públicas que este usuário é membro;
- **user\_favourites\_count**: quantidade de *tweets* que determinado usuário curtiu ao longo do tempo de vida de sua conta;
- **user\_statuses\_count**: quantidade de *tweets* incluindo *retweets* emitidos pelo usuário;
- **user\_created\_at**: data de criação da conta do usuário no Twitter;
- **user\_profile\_image\_url**: url da imagem de perfil do usuário;
- **user\_profile\_image\_url\_https**: endereço completo da imagem de perfil do usuário;
- **user\_following**: quantidade de contas que o usuário é seguidor;
- **created\_at**: informação referente a hora que o *tweet* foi criado;
- **id\_str**: identificador único do *tweet*;
- **text**: texto publicado pelo usuário;

- `source`: dispositivo de origem utilizado para a publicação do *tweet*;
- `in_reply_to_status_id_str`: caso o *tweet* seja uma resposta ao *tweet* original, este campo armazena o identificador do *tweet* original;
- `in_reply_to_screen_name`: armazena o apelido do autor original do *tweet*;
- `coordinates_lng`: coordenadas de longitude caso o dispositivo esteja com a localização GPS ativada na hora da publicação;
- `coordinates_lat`: coordenadas de latitude caso o dispositivo esteja com a localização GPS ativada na hora da publicação;
- `retweet_count`: quantidade de vezes que um determinado *tweet* foi compartilhado por outros usuários;
- `favorite_count`: quantidade de vezes que um determinado *tweet* foi curtido por outros usuários;
- `entities_hashtags`: representa todos as *hashtags* que foram analisadas foram analisadas no texto do *tweet*;
- `entities_user_mentions_id_str`: representa o nome dos usuários que foram mencionados no texto do *tweet*;
- `entities_user_mentions_screen_name`: representa o apelido dos usuários que foram mencionados no texto do *tweet*;
- `retweeted_status_userid`: contém uma representação do *tweet* original publicado;
- `retweeted_status_username`: contém uma representação do nome do usuário que publicou o *tweet* original;