XXV Workshop de Iniciação Científica da Embrapa Gado de Leite
Juiz de Fora, MG – 5 de março de 2020

Quais os tipos de queijos mais comentados no Twitter[1]

Nedson D. Soares[1], Emerson Campos[2], Thallys Nogueira[2], Kennya Siqueira[3,4], José Maria N. David, Regina Braga

[1]Este trabalho foi apresentado com o apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.
[2]Mestrando em Ciência da Computação– UFJF, Juiz de Fora, MG. e-mail: nedson@ice.ufjf.br.; jose.david@ufjf.edu.br
[2]Pesquisador - Embrapa Gado de Leite, Juiz de Fora, MG. e-mail: kennya.siqueira@embrapa.br emersonwendelim@gmail.com; thallysnogueira@ice.ufjf.br; regina.braga@ufjf.edu.br
[3]Orientador

Resumo: Redes sociais online como Facebook, Twitter e Instagram estão entre as maiores inovações da internet moderna. Por meio dessas redes, os usuários podem consumir e publicar dados. A capacidade de extrair informações dessa grande quantidade de dados é essencial para a sobrevivência e a modernização das empresas. Com esse objetivo, este trabalho apresenta uma arquitetura que combina técnicas de processamento de linguagem natural (PLN), mineração de dados e ontologias para analisar o conteúdo e a propagação de informações nas redes sociais, com aplicação no mercado lácteo brasileiro. As técnicas foram utilizadas para analisar mensagens postadas no Twitter relacionadas ao queijo com o propósito de identificar os principais tipos de queijo comentados nesta rede social. Como resultado, a metodologia traz a palavra coalho como a que mais se destacou no período analisado, e esta está relacionada ao queijo coalho. A ontologia desenvolvida mostrou resultado satisfatório na organização do conhecimento adquirido do domínio do queijo, e foi capaz de processar e apresentar informações sobre queijos.

Palavras–chave: mineração de dados, ontologia, processamento de linguagem natural, PLN, rede social online

Which types of cheese are most commented on Twitter

Abstract: Online social networks like Facebook, Twitter and Instagram are among the greatest innovations of the modern Internet. Through these networks, users can consume and publish data. The ability to extract information from this large amount of data is essential for the survival and the modernization of companies. With this purpose, this work presents an architecture that combines techniques of natural language processing (PLN), data mining and ontologies to analyze the content and the propagation of information on social networks, with application in the Brazilian dairy market. The techniques were used to analyze messages posted on Twitter related to cheese in order to identify the main types of cheese commented on this social network. As a result, the methodology brings the word “coalho” as the one that stood out the most in the analyzed period, and this is related to coalho cheese. The developed ontology showed a satisfactory result in the organization of the knowledge acquired in the cheese field, and was able to process and present information about cheeses.

Keywords: natural language process, mining data, ontology, online social network, OSN

Introdução

Conteúdos compartilhados em redes sociais tendem a demonstrar características associadas ao perfil de cada usuário, principalmente seus interesses e opiniões relacionadas a diferentes assuntos. Com redes sociais contendo uma quantidade significativa de usuários ativos e podendo ser acessadas de diversos dispositivos, uma grande quantidade e diversidade de conteúdo é produzido diariamente. Nesse cenário, o Twitter destaca-se como uma das maiores redes sociais da atualidade, possuindo mais de 300 milhões de usuários ativos mensalmente. O Brasil tem o segundo maior número de usuários da rede, logo atrás dos Estados Unidos, com mais de 27,7 milhões de contas ativas (Emarketer, 2016). O Twitter ainda disponibiliza gratuitamente uma *Application Programming Interface* (API) para mineração de dados públicos criados por seus usuários, facilitando a obtenção dessas informações.

Segundo Pak & Paroubek (2010) e Araújo *et al.* (2014), analisar os conteúdos compartilhados em redes sociais pode auxiliar no entendimento da opinião das pessoas sobre diferentes assuntos, e, a partir desta análise, dentre diversas outras aplicações possíveis, empresas podem saber mais sobre o que os seus consumidores pensam sobre seus produtos ou serviços (Pushpam & Jayanthi, 2017).

Assim, as redes sociais podem ser uma alternativa às pesquisas de mercado tradicionais, que, muitas vezes, são dispendiosas, demoradas e sem representatividade num país de dimensões continentais como o Brasil. Neste contexto, o presente trabalho visa analisar o mercado de queijos brasileiros por meio do Twitter. Os queijos estão em segundo lugar no ranking dos derivados lácteos mais consumidos no Brasil em 2017, e por isso se apresentam como uma boa oportunidade para análise de conteúdo. Em volume total de vendas, os queijos obtiveram crescimento de 124% no período de 2005 a 2016 (Siqueira, 2019). Assim, este estudo propõe o desenvolvimento de uma arquitetura que permita identificar os tipos de queijos mais comentados no Twitter, como forma de captar as preferências dos consumidores de queijo no Brasil.

Material e Métodos

Este trabalho apresenta a construção de uma arquitetura por meio: (i) da mineração de dados textuais que continham as palavras “queijo” e “#queijo” no Twitter; (ii) do processamento de linguagem natural dos textos extraídos; e (iii) da criação de uma ontologia para organização semântica do conhecimento no setor de produção de queijo.

Resultados e Discussão

O resultado do processo de coleta se deu em 82.868 *tweets* da língua portuguesa do Brasil durante o período de 10 dias. Por meio do pré-processamento do conteúdo, foram removidos 33.269 *tweets* duplicados e extraídos 3.590 *tweets* que citam, pelo menos uma vez, as palavras que correspondem aos tipos de queijos mais conhecidos.

Considerando o conjunto de palavras dos *tweets* extraídos, a Figura 1 apresenta um histograma contendo a distribuição de frequências das 10 palavras mais frequentes. Por meio desta, foi possível observar que a palavra correspondente ao tipo de queijo mais citado foi a palavra coalho, ficando à frente de requeijão e cheddar.

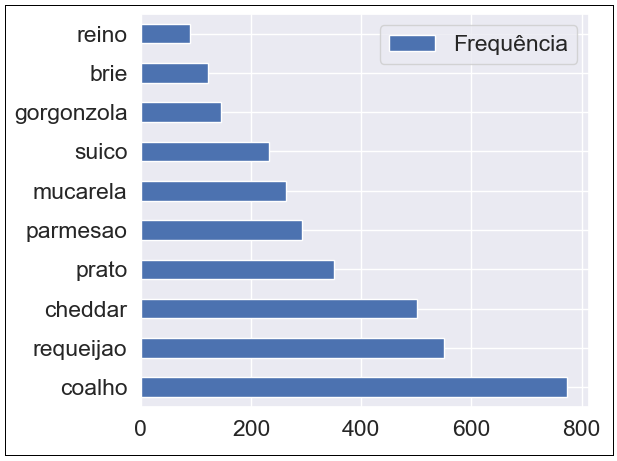

Figura 1. Histograma dos 10 tipos de queijo mais comentados no Twitter.

Torna-se importante ressaltar que a palavra coalho, não necessariamente refere-se ao queijo coalho. Ela pode se referir ao coalho usado na produção de queijo artesanal. No entanto, as palavras requeijão e cheddar referem-se aos tipos de queijos consumidos. Neste ponto, é interessante ressaltar que, apesar do curto período de tempo analisado, a pesquisa refletiu a realidade, já que o requeijão é, atualmente, o principal queijo do mercado de *commodities*, assumindo o lugar da muçarela a partir de 2012.

A ontologia desenvolvida reflete parte da árvore genealógica do leite, acrescida de informações sobre queijos mineiros e outros. Como a palavra coalho foi a que mais se destacou no período analisado, e esta está relacionada ao queijo coalho, que é um queijo artesanal, uma análise preliminar de ontologia foi realizada para Queijo Minas Artesanal (QMA), visto que este possui mais informações detalhadas do modo de fazer e características do que o queijo coalho. A Figura 2 apresenta uma visão sub expandida da classe "Queijo" na ontologia desenvolvida.

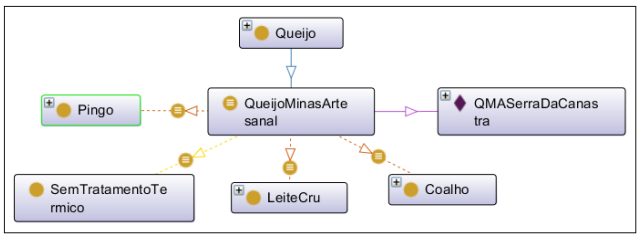

Figura 2. Visão sub expandida dos relacionamentos inferidos pela ontologia.

De acordo com a taxonomia do queijo organizada, a ontologia foi capaz de classificar indivíduos na classe "QueijoMinasArtesanal". Além disso, ela ainda infere relação com a classe "SemTratamentoTermico", que corresponde ao processo de produção do QMA. Um exemplo dessas relações pode ser observado na Figura 2, onde "QMASerraDaCanastra" é o indivíduo exemplo em questão.

Conclusões

A abordagem desenvolvida permitiu a coleta de *tweets* sobre o queijo em tempo real. Porém, para concluir o tipo de queijo mais consumido no Brasil, é necessária uma mineração de dados variada, ou seja, em diferentes repositórios e coletados em diferentes períodos do dia e

tempo. Contudo, este primeiro passo sugere que a arquitetura desenvolvida está em conformidade com a realidade.

A ontologia desenvolvida organiza o conhecimento baseado na árvore genealógica do leite em linguagem OWL. Trabalhos futuros devem focar no desenvolvimento de uma API para processar as informações dos queijos retiradas dos *tweets*.

Agradecimentos

Ao Programa Residência Zootécnica Digital da Embrapa Gado de Leite pela concessão da bolsa.

Referências

ARAÚJO, M., GONÇALVES, P., CHA, M., & BENEVENUTO, F. **iFeel.** Proceedings of the 23rd International Conference on World Wide Web - WWW 14 Companion. ACM Press, 2014;

EMARKETER. **Twitter's User Base to Grow by Double Digits This Year**. 22, julho 2016. Disponível em: <https://www.emarketer.com/Article/Twitters-User-Base-Grow-by-Double-Digits-This-Year/1014243>. Acesso em: 18 fev. 2020;

PAK, A., & PAROUBEK, P. **Twitter as a Corpus for Sentiment Analysis and Opinion Mining**. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 5 de 2010;

PUSHPAM, C. A., & JAYANTHI, J. G. **Overview on Data Mining in Social Media**. International Journal of Computer Sciences and Engineering, 5, 147-157, 11 de 2017;

REBALA, G., RAVI, A., & CHURIWALA, S. **Natural Language Processing. Em An Introduction to Machine Learning** (pg. 117-125). Springer International Publishing, 2019. doi:10.1007/978-3-030-15729-6_10;

SIQUEIRA, K. B. **O mercado consumidor de leite e derivados.** Juiz de Fora: Embrapa Gado de Leite - Circular Técnica (infoteca-e), 2019. 17 p. (Embrapa Gado de Leite. Circular Técnica, 120). Disponível em: <http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1110792>. Acesso em: 18 fev. 2020.