Uma Arquitetura para a Recomendação de Consumidores de Queijo Artesanal Brasileiro

Nedson D. Soares[1], Regina Braga[1], José Maria N. David[1], Kennya B. Siqueira[2], Victor Ströele[1], Fernanda Campos[1]

[1]Instituto de Ciências Exatas – Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora – MG – Brasil

[2]Embrapa Gado de Leite
Juiz de Fora – MG – Brasil

{nedson, victor.stroele}@ice.ufjf.br, {regina.braga, jose.david, fernanda.campos}@ufjf.edu.br, kennya.siqueira@embrapa.br

Abstract. *Extracting information from social networks has become essential for the survival and modernization of many companies. With this purpose, this work presents an architecture capable of searching, analyzing and recommending the content and the propagation of information on social networks, considering the Brazilian dairy market. Using ontologies and inference mechanisms, the architecture is capable of supporting the classification of user content and present it through visualization mechanisms. Through this architecture, we aimed to support market research conducted at Embrapa Gado de Leite. The results obtained in a feasibility study were satisfactory and demonstrated that the architecture provides support to researchers.*

Resumo. *Coletar informações de redes sociais se tornou essencial para a sobrevivência e modernização de muitas empresas. Com este objetivo, este trabalho apresenta uma arquitetura capaz de buscar, analisar e recomendar o conteúdo e a propagação de informações nas redes sociais, considerando o mercado de laticínios brasileiro. Usando ontologias e mecanismos de inferência, a arquitetura é capaz de suportar a classificação do conteúdo do usuário e apresentá-la através de mecanismos de visualização. Com essa arquitetura, buscamos apoiar pesquisas de mercado realizadas na Embrapa Gado de Leite. Os resultados obtidos em um estudo de viabilidade foram satisfatórios e mostraram que a arquitetura fornece suporte aos pesquisadores.*

1. Introdução

Ferramentas inteligentes são cada vez mais necessárias na indústria moderna, permitindo inovações mais eficazes [Galleta *et al*. 2018]. Ao tratarem grandes volumes de dados essas ferramentas têm utilizado tecnologias de *big data* [McAfee *et al.* 2012]. A análise destes dados permite a descoberta de correlações e, com isso, a derivação de novo conhecimento. Por exemplo, as informações geradas a partir do uso de ontologias [Guarino 1998]. Com isso, existe uma mudança na forma em que os dados são analisados, por meio do processamento de inferências na ontologia, que permite a descoberta de novos relacionamentos entre os dados.

Redes Sociais Online (OSN, do inglês *Online Social Network*) estão entre as maiores plataformas da Internet moderna. Elas possuem uma quantidade significativa de usuários e podem ser acessadas através de diversos dispositivos. Com o intuito de conectar pessoas e conteúdos compartilhados, as OSNs apresentam características associadas ao perfil de cada usuário, principalmente com relação a interesses e opiniões em diferentes assuntos [Pak e Paroubek 2010].

Uma grande quantidade de dados, de diversos tipos, é produzida diariamente pelas OSNs, tornando-as um meio promissor para a coleta de dados relacionados a opiniões e hábitos de cada usuário. Muitas vezes, os conteúdos postados nas OSNs possuem *links* para serviços, eventos, produtos, pessoas, dentre outros. Segundo Pak e Paroubek (2010), a análise desses conteúdos compartilhados pode auxiliar no entendimento da opinião das pessoas sobre diferentes assuntos. Neste sentido, consideramos que as OSNs podem contribuir na análise de estruturas do agronegócio. Usando seus dados sociais como principais elementos de análise, as OSNs possibilitam a análise de estruturas produtivas no setor de agronegócios [Talamini e Ferreira 2010].

Dentro do agronegócio brasileiro, um setor que se destaca é o de leite e derivados. A indústria de laticínios brasileira é um dos setores mais importantes da indústria alimentos no País, perdendo em faturamento apenas para a indústria de carnes. E um derivado lácteo que tem apresentado aumento significativo de consumo no Brasil, são os queijos, em especial, os queijos artesanais.

O País possui diversas regiões produtoras de queijo artesanal que cultivam tradições seculares na fabricação desses queijos. No entanto, o setor carece de pesquisas sobre o mercado consumidor desses produtos.

Considerando diferentes características de perfil de consumidor e os desafios do setor lácteo, as pesquisas de mercado tradicional demandam tempo, são caras e, por vezes, incompletas e sem representatividade. Neste contexto, estudos realizados na Embrapa Gado de Leite [Nardy *et al.* 2019] apontam oportunidades para novos ambientes que poderiam proporcionar maiores margens no mercado, mas que tais caminhos não são os tradicionais. Assim, propomos um ambiente inteligente para análise de dados baseado em conteúdo de OSNs, para extrair, processar e permitir a recomendação do resultado como uma inovação de impacto para o segmento de produção de queijo artesanal. Neste sentido, este estudo propõe uma arquitetura para apoiar as pesquisas de mercado no segmento de laticínios, mais especificamente na produção de queijos artesanais.

A principal contribuição deste trabalho é a especificação de uma arquitetura de recomendação baseada na identificação do perfil dos consumidores do setor lácteo para auxiliar a tomada de decisão. O artigo está organizado da seguinte forma: Seção 2 detalha o referencial teórico da pesquisa. Seção 3 apresenta os trabalhos relacionados a busca e recomendação de dados. Seção 4 apresenta uma arquitetura de recomendação baseada em dados de OSNs. Seção 5 descreve um estudo de viabilidade da arquitetura considerando o domínio de queijos artesanais. Finalmente, a Seção 6 apresenta as considerações finais.

2. Referencial Teórico

Big data é um conceito que se refere à quantidade significativa de dados gerados a todo momento por diversas fontes, como sensores industriais, OSNs ou documentos da Web [McAfee *et al.* 2012]. Para prospectar informações relevantes, estes dados necessitam de técnicas de análise de *big data*. Nas OSNs, os usuários se dividem em diversos grupos

sociais, os quais tendem a compartilhar e difundir informações em aspectos específicos do grupo. Por exemplo, produtores e consumidores de queijos artesanais podem formar um grupo e os tópicos discutidos por eles tendem a ser conteúdos relacionados aos derivados do queijo. Baseado nisso, conceitos utilizados em sistemas de recomendação podem ser utilizados para lidar com a recomendação dos usuários [Yang *et al.* 2020]. Assim, um usuário na OSN pode ser um possível consumidor de queijo artesanal. Sua informação histórica pode ser utilizada como registro de compra e as informações contidas nos textos podem ser consideradas como recursos de produto. Como resultado, o sistema julga se existe relação entre os recursos e usuários, e os recomenda como potencial consumidor.

Neste contexto, os sistemas de recomendação utilizam mecanismos que ajudam a encontrar relações entre usuários e produtos de interesse [Batmaz *et al.* 2018]. Estes sistemas são tradicionalmente classificados por autores como *(i)* colaborativos; *(ii)* baseados em conteúdo; *(iii)* híbridos; *(iv)* demográficos e *(v)* baseados em conhecimento [Aggarwal 2016]. Neste artigo, utilizamos apenas as classificações *(iv)* onde a recomendação é baseada em parâmetros de localização e *(v)* que sugere itens baseados em inferências sobre as características do usuário (ex. ontologias). As demais classificações não são adequadas, pois dependem de parâmetros que não são tratados neste artigo para fazer a recomendação dos consumidores.

Para classificar o conteúdo textual extraído das OSNs é essencial obter um entendimento preciso da semântica do texto [Abu-Salih 2018]. O uso de ontologias [Guarino 1998] pode auxiliar nessa tarefa, conforme apresentamos na seção 4.

3. Trabalhos relacionados

Alguns artigos abordam os principais conceitos de sistemas de recomendação. Abel *et al.* (2011) compararam perfis de usuário, baseados em recursos de texto e *hashtags* em *tweets*, para recomendar notícias. Porém, os resultados obtidos por meio de uma avaliação mostraram a necessidade do uso de semântica para aumentar a precisão das análises ao inferir conhecimento implícito. Chianese *et al.* (2016) fizeram um estudo baseado em ontologia para identificar indicadores de desempenho do patrimônio cultural expressos pelos usuários de OSN. De Nart *et al.* (2016) propuseram uma extração de tópicos de *tweets* baseada em conteúdo utilizando aprendizado de máquina. Tao *et al.* (2012) apresentaram o TUMS, uma arquitetura baseada em serviços para inferir perfis de usuários semânticos a partir de *tweets*.

Outros estudos apontam os benefícios de buscar e recomendar em tempo real especialistas que estão entre usuários de OSNs [Yang *et al.* 2020]. Devido a sua popularização, os dados relacionais encontrados nas OSNs são de grande importância para estimar o grau de associação entre a pessoa e o tópico. Porém, alguns estudos [Chen *et al.* 2016] não consideram o fator temporal. O conhecimento dos usuários evolui ao longo do tempo e seu interesse pode ser alterado. Portanto, para inferir tópicos reais de interesse é necessária uma análise ao longo do tempo [Abu-Salih *et al.* 2018].

Comparados às pesquisas previamente apresentadas, o nosso trabalho estima o grau de relação entre usuário e tópico em tempo real pela quantidade de documentos relacionados ao usuário em questão. Além disso, os sistemas de recomendação apresentados não são focados no mercado de queijos artesanais e possuem especificidades que não atendem à arquitetura de recomendação proposta. Portanto, foi desenvolvida uma

arquitetura específica para o setor de produção de queijos artesanais, na qual a análise de redes sociais foi utilizada para identificar consumidores desses produtos e recomendá-los.

4. Arquitetura de Recomendação de Consumidores de Queijos Artesanais

Para identificar e sugerir consumidores e colaborar no método de fazer pesquisa de mercado relacionada ao mercado lácteo, foi proposta uma arquitetura de recomendação de consumidores de produtos lácteos, especificamente para o mercado de queijos artesanais.

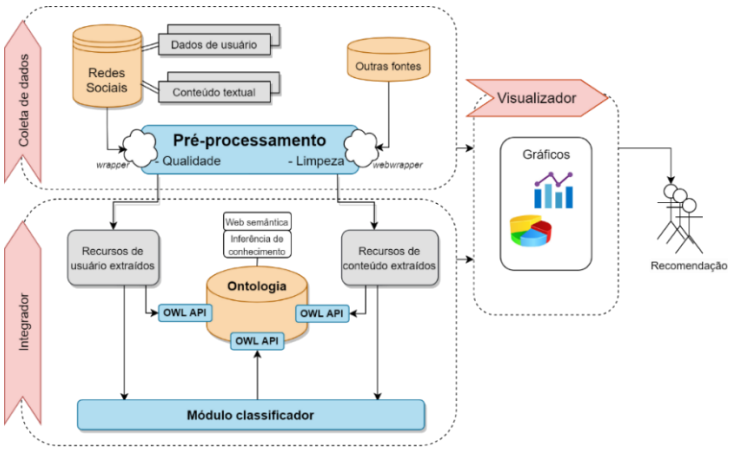

Figura 1. Arquitetura proposta.

O processamento da arquitetura (Figura 1) é composto por 3 subsistemas, que são descritos nas seguintes etapas: *(i)* informações sobre o setor lácteo, em especial sobre o queijo artesanal, são coletadas das OSNs e outras fontes Web relacionadas, e são tratadas utilizando processamento de texto [Young *et al*. 2018] para dar qualidade aos dados extraindo apenas o que é relevante à pesquisa; *(ii)* os componentes da ontologia e um módulo classificador processam estas informações, analisando e inferindo (por meio de algoritmos de inferência da ontologia) os recursos extraídos; por fim *(iii)* a recomendação é disponibilizada através do componente de visualização. Essas etapas são descritas com mais detalhes nas subseções abaixo.

4.1. Coleta de dados

As informações de OSN são obtidas por meio de um *wrapper* que contém uma API para a mineração de dados nos bancos oficiais da OSN. Neste artigo, empregou-se a API do Twitter que utiliza um conjunto de restrições e palavras-chaves relacionadas ao queijo artesanal para buscar conteúdos que sejam relevantes à pesquisa. Em conjunto a este *módulo*, um *webwrapper* também é utilizado para coletar dados da web consultando outras fontes, como por exemplo o site oficial do Instituto Mineiro de Agropecuária (IMA), que contém documentos importantes sobre a produção de queijo. As restrições de busca neste módulo são as mesmas do *wrapper*. Ainda nessa etapa, os dados coletados são pré-processados utilizando a biblioteca de código aberto para processamento de texto: NLTK[1]. Essa biblioteca faz uso de técnicas que podem sofrer muitas interferências por conta dos ruídos. Ela possui algoritmos estáticos que utilizam a frequência como

[1] http://www.nltk.org/book/ch00-pt.html

indicador. Nesses algoritmos, os ruídos são palavras que aparecem muitas vezes, porém não têm significado expressivo para o contexto do texto. Por exemplo, no *tweet* "*eu gosto de queijo coalho!* " as palavras "eu" e "de" são consideradas ruídos pela aplicação e são removidas para gerar qualidade ao conjunto de dados de palavras da arquitetura.

4.2. Integrador

Neste subsistema, a extração de recursos e a inferência de conhecimento são utilizados para extrair informações, podendo ser dividido em duas etapas: *(i)* análise do conjunto de palavras pertencentes ao usuário; e *(ii)* análise das informações individuais dos perfis coletados.

4.2.1. Análise de Conteúdo

Para analisar o conteúdo desenvolvemos um algoritmo em Python: O *módulo* classificador (Figura 1). Ele é responsável por enriquecer e classificar os dados por meio de algoritmos de classificação e aprendizado de máquina. Utilizamos neste *módulo* a biblioteca da Azure Microsoft[2] e seus serviços cognitivos gratuitos para *(i)* analisar os recursos de usuário (ex. fotos de perfil) para estimar informações como idade e gênero; e *(ii)* polarizar suas mensagens (ex. neutro, positivo e negativo) e auxiliar na descoberta de opinião dos consumidores. As informações para análise de falsos positivos (nos quais uma mensagem foi polarizada incorretamente) são limitadas aos serviços grátis da Azure. Porém, a biblioteca disponibiliza a porcentagem de acurácia do resultado da polarização de cada mensagem. Assim, essa informação é atribuída junto à polarização da mensagem.

Além disso, o *módulo* ainda conta com uma função (F1) simples que desenvolvemos para rotular o conteúdo textual de acordo com os tipos de queijos citados mais conhecidos. Os dados classificados são utilizados para enriquecer a ontologia. Um exemplo destes dados pode ser visualizado na Tabela 1, a qual contém uma lista em ordem decrescente dos cinco usuários que mais postaram conteúdo dentro do conjunto de dados coletados.

Usuário	Gênero	Idade	Tipos de Queijo	Sentimentos
1	masculino	33	reino[2], **requeijão[6]**	negativo[2], **positivo[4],** neutro[2]
2	feminino	31	prato[1], reino[1], **coalho[4],** muçarela[1]	negativo[2], positivo[2], **neutro[3]**
3	masculino	33	prato[1], **cheddar[3],** requeijão[2]	negativo[2], **positivo[3],** neutro[1]
4	feminino	22	parmesão[2], **brie[4]**	negativo[3], positivo[3]
5	feminino	13	**reino[2],** prato[1], **coalho[2],** muçarela[1]	negativo[1], **positivo[3], neutro[2]**

Tabela 1. Usuários classificados.

4.2.2. Análise Ontológica

Para organizar e inferir conhecimento, propomos o uso de uma ontologia [Guarino 1998]. Assim, foi especificada uma ontologia a partir da Árvore Genealógica do Leite[3] de modo a organizar e extrair informações relevantes, utilizando regras semânticas. Para popular a ontologia, implementamos um *wrapper* contendo a OWL API da linguagem Python: OwlReady2[4]. Os recursos de conteúdo e usuário são utilizados pelo *wrapper* para criar

[2] https://pypi.org/project/azure/

[3] https://www.arvoredoleite.org/

[4] https://pythonhosted.org/Owlready2/

novas instâncias e gerar conhecimento na ontologia. Identificamos classes, *data properties*, *object properties*, etc. A Figura 2 apresenta uma visão geral dessa ontologia com foco no indivíduo "consumidor1" por meio de um gráfico.

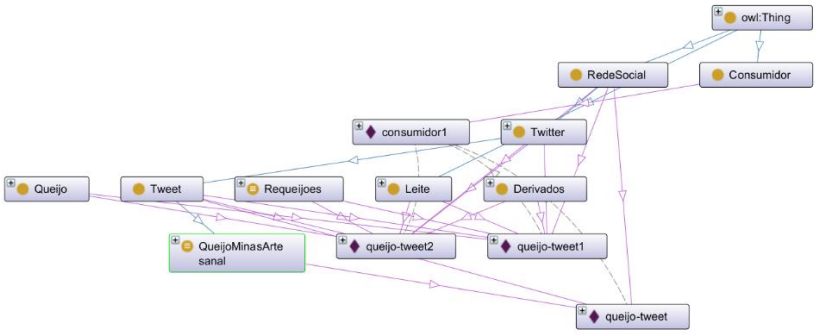

Figura 2. Visão geral das principais classes da ontologia.

As regras da ontologia foram desenvolvidas utilizando linguagem SWRL[5] e são apresentadas na Figura 3. Para testar a ontologia, criamos o indivíduo "consumidor1" que citou em seu *tweet* as palavras "coalho", "leite cru", "pingo". De acordo com a taxonomia do queijo organizada, a ontologia foi capaz de classificar o indivíduo como consumidor de um queijo do tipo "QueijoMinasArtesanal". A ontologia pode ser acessada em [6].

```
"Tweet(?tw)^Consumidor(?co)^usuario(?tw,?us1)^usuario(?co,?us2)^swrlb:equal(?us1, ?us2)->tuitadoPor(?tw,?co)"
"municipio(?cid,?mu1)^municipio(?tw,?mu2)^swrlb:equal(?mu1,?mu2)^Cidade(?cid)->tuitadoEm(?tw,?cid)"
"produto(?x, ?a)^nome(?y, ?b)^swrlb:equal(?a, ?b)->temProduto(?x,?y)"
"Produtor(?pr)^municipio(?cid,?mu1)^municipio(?pr, ?mu2)^swrlb:equal(?mu1, ?mu2)->produzEm(?pr,?cid)"
```

Figura 3. Regras SWRL.

5. Estudo de Viabilidade

Para verificar a viabilidade técnica da arquitetura proposta e, então, auxiliar na avaliação reunindo informações sobre elementos da arquitetura no contexto de aplicação, foi definido o escopo da avaliação através do GQM (do inglês *Goal*, *Questions*, *Metrics*): "**Analisar** o uso da arquitetura de recomendação do **ponto de vista** de pesquisadores e produtores lácteos **no contexto** do grupo de pesquisa da Embrapa Gado de Leite que têm como foco os consumidores de queijo artesanal. Para orientar este estudo, a questão de pesquisa principal é: **RQ. Quem são os potenciais consumidores de queijo artesanal no Brasil?**

Guiados pela RQ, coletamos uma lista de usuários do Twitter e seu conteúdo utilizando palavras-chave relacionadas ao queijo, conforme discutido na Seção 4.1. Utilizando um filtro de palavras chaves relacionados ao queijo artesanal, o resultado do *wrapper* foi de 412.865 *tweets* da língua portuguesa do Brasil, tuitados em diferentes partes do mundo durante os períodos do ano de 2020: *(i)* de 11/fev. até 23/mar.; e *(ii)* de 01/abr. até 06/abr. Para evitar falsos positivos, que neste caso se trata dos *tweets* que falam de queijo mas não citam tipo nenhum, apenas os *tweets* classificados por F1 são extraídos para a análise. O *webwrapper* extraiu documentos com informações sobre queijos artesanais do site oficial do IMA (Instituto Mineiro de Agropecuária). O processo resultou

[5] https://www.w3.org/Submission/SWRL/

[6] https://github.com/nedsons/ontologia-oc

em documentos contendo as legislações, produtores cadastrados e documentos para registro. Estes documentos foram utilizados para gerar conhecimento na ontologia.

5.1. Resultados

Por meio do pré-processamento do conteúdo coletado, a análise dos dados permitiu verificar a viabilidade da arquitetura no suporte às pesquisas relacionadas a queijo artesanal no contexto da Embrapa Gado de Leite, a partir das respostas as questões de pesquisa propostas. Foram removidos 158.849 *tweets* duplicados e extraídos 11.845.

RQ. **Quem são os potenciais consumidores de queijo no Brasil?** A arquitetura pode apoiar as investigações e avaliações para a indústria, governo e pesquisadores do setor lácteo proporcionando uma estrutura padronizada e organizada. Com isso, eles poderão ser mais ágeis em seus estudos. Além disso, as visualizações dos dados classificados e agrupados podem auxiliar os pesquisadores na identificação dos consumidores. A arquitetura rotulou 10.403 usuários de acordo com sua idade, gênero, sentimento e tipo de queijo. Com isso, os usuários obtidos podem ser considerados potenciais candidatos a consumidores auxiliando na pesquisa de mercado de queijos.

Por meio de consultas à ontologia da arquitetura, a solução proposta simplifica a recomendação de consumidores de queijo para os pesquisadores que necessitam saber sobre um determinado perfil de usuário, bem como para definir um novo perfil. Através das regras de inferência na ontologia, a arquitetura rotula os consumidores, fornecendo ao pesquisador uma maneira mais acessível de buscar informações. Por exemplo, quando um pesquisador precisa saber a faixa etária de consumidores do queijo coalho para traçar novas estratégias de marketing do produto em análise, ou então precisa saber qual o tipo de queijo mais consumido pelas mulheres. Por fim, a recomendação de consumidores na arquitetura se dá por meio da visualização dos dados e consultas à ontologia.

6. Considerações finais

Este artigo apresentou uma arquitetura de recomendação de consumidores relacionados ao mercado de queijo artesanal, baseado em dados de OSNs. Um estudo de viabilidade técnica dessa arquitetura foi realizado para que pudéssemos ter informações sobre a arquitetura. Essas informações podem apoiar na tomada de decisões em relação a mudanças na estrutura da arquitetura, evitando problemas futuros. Como trabalhos futuros, propõe-se explorar outros domínios e as técnicas de aprendizado de máquina para apoiar as recomendações. Uma análise em outros domínios do conhecimento será conduzida para obter uma visão abrangente da área, facilitando o desenvolvimento de ontologias e viabilizando a arquitetura em outro contexto. Diferentes algoritmos de classificação de aprendizado de máquina serão também explorados. Assim, a arquitetura poderá prever com mais eficiência a probabilidade de interesse entre usuário e tópico no contexto de aplicação. Conteúdo multimídia e as informações geográficas também serão exploradas para definir uma segmentação inteligente do usuário. Além disso, cabe realizar uma validação mais robusta, com grandes volumes de dados, para verificar a escalabilidade da abordagem proposta.

Agradecimentos: Este trabalho foi parcialmente financiado pela UFJF, CAPES, CNPq, Programa Residência Zootécnica Digital da Embrapa Gado de Leite e a Rede Integrada de Pesquisa em Alta Velocidade (RePesq) da UFJF.

Referências

Abel F., Gao Q., Houben GJ., Tao K. (2011) “Analyzing User Modeling on Twitter for Personalized News Recommendations”, UMAP 2011, Lecture Notes in CS, p. 1-12.

Abu-Salih, B., Wongthongtham, P., & Chan, K. Y. (2018) “Twitter mining for ontology-based domain discovery incorporating machine learning”, Journal of Knowledge Management, p. 949-981.

Aggarwal, C. C. (2016) “An Introduction to Recommender Systems Recommender”, Springer International Publishing, p. 1-28.

Batmaz, Z.; Yurekli, A.; Bilge, A. & Kaleli, C. (2018) “A review on deep learning for recommender systems: challenges and remedies”, A.I. Review, p. 1-37.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016) “Topic modeling for evaluating students’ reflective writing: a case study of pre-service teachers’ journals”, Proceedings of the Sixth International Conference on LAK’16, ACM, p. 1-5.

Chianese, A.; Marulli, F. & Piccialli, F. (2016) “Cultural Heritage and Social Pulse: A Semantic Approach for CH Sensitivity Discovery in Social Media Data”, IEEE ICSC, p. 1-6.

Galletta, A., Carnevale, L., Celesti, A., Fazio, M., & Villari, M. (2018) “A Cloud-Based System for Improving Retention Marketing Loyalty Programs in Industry 4.0: A Study on Big Data Storage Implications”, IEEE Access*, 6*, 5485–5492.

Guarino, N. (1998) “Formal ontology in information systems”, Proceedings of the first international conference (FOIS'98), Trento, Italy IOS press, p. 3-15.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012) “Big data: the management revolution”, Harvard business review, p. 60–68.

Nardy, V. P. D. R., Carvalho, G. R. & da Rocha, D. T. (2019) “Mercado de leite fluido e queijos no Brasil: uma análise de 2005 a 2016”, Embrapa Gado de Leite, p. 1-5.

Nart, D. D., Degl'Innocenti, D., Basaldella, M., Agosti, M. & Tasso, C., (2016) “A Content-Based Approach to Social Network Analysis: A Case Study on Research Communities”, CCIS, Springer International Publishing, p. 142-154.

Pak, A., & Paroubek, P. (2010) “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, Proceedings of the 7th International Conference on LREC'10, p. 1320-1326.

Siqueira, K. B. (2019) “O mercado consumidor de leite e derivados”, Embrapa Gado de Leite - Circular Técnica (infoteca-e), p. 1-17.

Talamini, E. & Ferreira, G. M. V. (2010) “Merging netchain and social network: Introducing the social netchain concept as an analytical framework in the agribusiness sector”, African journal of business management, Academic Journals, p. 2981-2993.

Tao, K., Abel, F., Gao, Q. & Houben, G.-J. (2012) “TUMS: Twitter-Based User Modeling Service Lecture Notes in Computer Science”, Springer Berlin Heidelberg, p. 1-15.

Yang, X., Dong, M., Chen, X., & Ota, K. (2020) “Recommender System-Based Diffusion Inferring for Open Social Networks”, IEEE TCSS, p. 24–34.

Young, T., Hazarika, D., Poria, S. & Cambria, E. (2018) “Recent Trends in Deep Learning Based Natural Language Processing”, IEEE CIM, p. 55-75.