UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Nedson Donato Soares

An Approach to foster Agribusiness through Data Analysis in Social Networks

Juiz de Fora
2023

Nedson Donato Soares

An Approach to foster Agribusiness through Data Analysis in Social Networks

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação. Área de concentração: Sistemas e Tecnologias da Computação.

Orientador: Prof. D.Sc. Regina Maria Maciel Braga

Coorientador: Prof. D.Sc. José Maria Nazar David

Juiz de Fora
2023

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Soares, Nedson Donato.
An Approach to foster Agribusiness through Data Analysis in Social Networks / Nedson Donato Soares. -- 2023.
111 f. : il.

Orientadora: Regina Maria Maciel Braga
Coorientador: José Maria Nazar David
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2023.

1. data analysis. 2. social networks. 3. dairy derivatives. 4. architecture. 5. ontology. I. Braga, Regina Maria Maciel, orient. II. David , José Maria Nazar, coorient. III. Título.

Nedson Donato Soares

An Approach to foster Agribusiness through Data Analysis in Social Networks

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Ciência da Computação. Área de concentração: Ciência da Computação.

Aprovada em 19 de janeiro de 2023.

BANCA EXAMINADORA

Profª. Dra. Regina Maria Maciel Braga - Orientadora
Universidade Federal de Juiz de Fora

Prof. Dr. José Maria Nazar David - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Victor Ströele de Andrade Menezes
Universidade Federal de Juiz de Fora

Prof. Dr. Márcio de Oliveira Barros
Universidade Federal do Estado do Rio de Janeiro

Dra. Kennya Beatriz Siqueira

Empresa Brasileira de Pesquisa Agropecuária

Juiz de Fora, 13/01/2023.

Documento assinado eletronicamente por **Victor Stroele de Andrade Menezes**, **Professor(a)**, em 20/01/2023, às 09:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jose Maria Nazar David**, **Professor(a)**, em 20/01/2023, às 10:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Regina Maria Maciel Braga Villela**, **Professor(a)**, em 20/01/2023, às 12:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Kennya Beatriz Siqueira**, **Usuário Externo**, em 20/01/2023, às 13:02, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcio de Oliveira Barros**, **Usuário Externo**, em 23/01/2023, às 07:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1110943** e o código CRC **24D9247A**.

Dedico este trabalho aos meus pais e irmãos que me inspiram e me auxiliaram na realização.

AGRADECIMENTOS

Agradeço a oportunidade de conclusão desta etapa a todos os professores do PPGCC, em especial aos meus orientadores Regina Braga e José Maria, por me proporcionar o conhecimento não apenas racional, mas a manifestação do caráter e afetividade da educação no processo de formação acadêmica, por tanto que se dedicaram a mim, não somente por terem me ensinado, mas por terem me feito aprender. A palavra mestre, nunca fará justiça aos professores dedicados aos quais sem nominar terão os meus eternos agradecimentos. Quero também agradecer a minha orientadora do projeto Residência Zootécnica Digital da Embrapa, Kennya Siqueira, que muito me ensinou sobre como trabalhar em grupo, e a todos aqueles envolvidos na residência com a Embrapa, cujos esforços e auxílio tornaram possível a concretização deste projeto. Agradeço também aos professores Márcio Barros e Victor Ströele por fazerem parte desta viagem como integrantes da banca de defesa da dissertação.

Aos amigos, companheiros de mestrado e irmãos na amizade que fizeram parte da minha formação e que vão continuar presentes em minha vida, com certeza. Agradeço ao meu grande amigo Tales, herói que me deu apoio, incentivo nas horas mais difíceis, de desânimo e cansaço. Aos meus pais e a Deus, pois sem eles nada disso seria possível. Obrigado.

“Mas para que o produto de uma pesquisa científica possa ser publicado não basta que ele apresente um conteúdo de qualidade, também é exigida qualidade de forma.” (MARÇAL JUNIOR, 2013, p. 19-20).

> Como definir uma biblioteca sem limitá-la a acervos e objetos? Se devo esperar que a biblioteca seja mais que um depósito de livros, o que devo esperar? O que faz uma biblioteca? Biblioteca como facilitadora. Em uma só palavra, o que biblioteca e bibliotecários fazem é facilitar. (LANKES, 2016, p.69)

RESUMO

O avanço da tecnologia da informação faz com que as redes sociais ganhem cada vez mais popularidade e inserção no cotidiano. Assim, a análise das opiniões e hábitos das pessoas é essencial para a modernização e sobrevivência das empresas. Nas redes sociais, as pessoas compartilham suas opiniões e acessam opiniões de outras pessoas sobre produtos, novidades e tendências, dando origem ao conceito de "pessoa influente". Uma pessoa influente (ou influenciador de mídia social) hoje é considerada uma estratégia de marketing. Por outro lado, a indústria de laticínios brasileira vem se destacando a cada ano, e uma das áreas promissoras é a produção de queijos. O mercado de queijos está em crescimento e pode atingir muitos consumidores. Com o objetivo de coletar informações das redes sociais para encontrar pessoas influentes, que apreciam queijos e que possam influenciar novos potenciais consumidores, este trabalho apresenta a IntelDigitalMarketing, uma proposta que engloba análise de redes sociais, recomendação e propagação de conteúdo, considerando o mercado brasileiro de queijos. Utilizando a IntelDigitalMarketing é possível identificar influenciadores e comunidades de usuários que abordam assuntos relacionados a temas específicos em informações de OSNs em diferentes redes sociais. Utilizamos a metodologia Design Science Research para verificar a viabilidade da proposta. A nossa solução é inovadora na medida em que engloba técnicas como redes complexas, machine learning e ontologias, para detectar tendências de mercado. Com essas técnicas combinadas, podemos detectar novos relacionamentos relevantes entre usuários que não são detectados por outras soluções semelhantes. Além disso, a solução proposta é online e em tempo real, tornando mais fácil acompanhar as tendências nas redes sociais. Os resultados mostraram que a solução permite a busca por comunidades de influenciadores digitais que falam sobre queijo, conhecimento sobre o que falam e a divulgação de informações na rede.

Palavras-chave: análise de dados, redes sociais, derivados do leite .

ABSTRACT

The advancement of information technology makes social networks gain more popularity and insertion in everyday life. Thus, the analysis of people's opinions and habits is essential for the modernization and survival of companies and institutions. In social networks, people share their opinions and access other people's opinions about products, news, and trends, giving rise to the concept of an "influential person". Today, an influential person (or social media influencer) is considered a marketing strategy. Considering specific markets, the Brazilian dairy industry has been standing out each year, and one of the promising areas is cheese production. The cheese market is growing and can reach many consumers. With the aim of collecting information from social networks to find influential people, who appreciate cheese and can influence new potential consumers, this work presents IntelDigitalMarketing. This proposal encompasses analysis of social networks, recommendation, and propagation of content, considering the Brazilian market of cheeses. Through its use, it is possible to identify influencers and user communities that address issues related to specific domains in different social networks. We used the Design Science Research methodology to verify the feasibility of the proposal. Our solution is innovative as it encompasses techniques such as complex networks, machine learning, and ontologies, to detect market trends. With these techniques combined, we can detect new relevant relationships between users that are not detected by other similar solutions. In addition, the proposed solution is online and in real-time, making it easier to follow trends in social networks. The results showed that the solution allows the search for communities of digital influencers who talk about cheese, what they talk about, and the dissemination of information on the network.

Keywords: data analysis, social networks, dairy derivatives.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABLV	Associação Brasileira da Indústria de Lácteos Longa Vida
API	Application Programming Interface
AQ	Assessment Questions
AUC	Area Under Curve
BPM	Business Process Management
BS	Backward Snowballing
CQ	Competency Questions
DSR	Design Science Research
FR	Functional Requirements
FS	Foward Snowballing
GQM	Goal, Questions, and Metrics
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NLP	Natural Language Processing
OSN	Online Social Network
OWL	Web Ontology Language
PICOC	Population, Intervention, Comparison, Outcome and Context
RQ	Research Question
SLM	Systematic Literature Mapping
SLMRQ	Systematic Literature Mapping Research Question
SMA	Social Media Analysis
SNA	Social Network Analysis
SPARQL	Protocol and RDF Query Language
SRQ	Secondary Research Questions
SWRL	Semantic Web Rule Language
UGC	User Generated Content

SUMÁRIO

1 INTRODUCTION

In this chapter we discuss this work's context, motivation and objectives. We also present the research question and the organization of the text.

1.1 CONTEXTUALIZATION

Nowadays, advances in information technologies turn different Online Social Networks (OSN), such as YouTube, Facebook, Instagram and Twitter, increasingly popular among Internet users (1). Due to this popularity, OSNs are becoming a crucial channel for business and marketing (2).

Understanding people through Social Network Analysis (SNA) plays an important role in analyzing marketing trends. The new knowledge extracted from SNA has many applications and generates value for organizations. According to (1), interest in studying human social networks has been explored more recently with the emergence of OSNs. The OSN provides access to voluntary information that was previously unattainable and demanded a high access cost.

From a structural point of view, OSNs are complex structures composed of vertices, usually representing users, and edges expressing some form of interaction, collaboration, or influence between these users, allowing the sharing and dissemination of information (3). Several companies have produced tools to detect leading users who exert some influence over other users and who are most likely to disseminate information about their products (4). This detection is necessary for the recent trend of investing in "word of mouth" marketing, where companies need to find out who are the people who will spread the information about the products. SNA can reveal information related to users and their interactions. As a result, it provides human behavior knowledge related to specific contexts, such as consumption profiles, and political trends, among other contexts (5,6).

Specifying models to correlate this data and identify specific users who follow specific patterns are necessary for these analyses. In this sense, OSNs can contribute to market surveillance in Agribusiness. Using social data as the main elements of analysis, OSNs can promote the analysis of productive structures in the Agribusiness sector. Among the main analysis types, we can highlight (i) the detection of influential users, who are individuals with significant impact on their OSNs and who can influence others; and (ii) the detection of communities, which is the discovery of groups with similar characteristics in the network (7). The analysis of the data generated by the OSNs can allow the discovery of implicit relationships, generating new information and allowing a new analysis cycle.

1.2 MOTIVATION

One sector that stands out in Agribusiness is milk and dairy products. The dairy industry is one of the most important sectors of the food industry, losing revenue only to the meat industry (8). In Brazil, the dairy industry represents a strategic economic sector, demanding huge investments. In the same trend, dairy derivatives have shown a significant increase in consumption in Brazil, notably cheese (9). Brazil has several regions producing cheese that maintain centuries-old traditions in the production.

According to ABLV (10), between 2017 and 2021 the dairy industry's milk volume increased. However, high temperatures and scarcity of rain caused a decrease from 2021 onwards. Therefore, the performance of this market dropped in 2021, which requires efficient strategies for a new growth cycle. However, the cheese was an exception and grew by 1.1%. Therefore, considering its importance, we investigate the cheese market as a starting point, considering it has a growing market and consequently can reach more consumers.

On the other hand, in today's business world, the large volume of data that needs to be analyzed for decision-making is driving companies to become increasingly dependent on investments in data infrastructure and in the analysis and use of data. Thus, there is a need for business intelligence to promote innovation as a process of collecting, organizing and analyzing information. The combination of intelligent techniques applied to the business can offer a competitive advantage for companies through the storage, analysis and visualization of big data using shared infrastructures.

1.3 PROBLEM

Agribusiness is one of the economic sectors that are strategic worldwide. Its growth is critical and can bring benefits to the world population. However, according to (11), Agribusiness still needs solutions that improve consumer market analysis (ex: product adequacy, sales, and cost reduction). Therefore, data analysis techniques can bring improvements to the sector. Traditional market research is time-consuming, expensive, and, at times, incomplete and without representation, considering the different characteristics of the consumer profile and the challenges of the Agribusiness sector. Dealing with multiple stakeholders in market research projects often makes quality data collection and communication reliable, low-cost, complex, and challenging. Studies carried out at Embrapa Gado de Leite (8,11) point to new opportunities that could improve the market. For example, the portal Observatório do Consumidor[1] which was created to monitor the consumer profile and consumption trends of milk and dairy products in Brazil (74).

The ability of organizations to react to changes imposed by the market is directly related to the absorption of information and generation of knowledge applied to the processes. In this vein, modern industry increasingly needs smart tools to generate new knowledge (12). These needs also apply to Agribusiness. Understanding consumer trends, perceptions, or preferences economically and efficiently is the focus of consumer science and manufacturing industries (13). When processing large volumes of data, intelligent tools have been using big data analytics, allowing the discovery of correlations and the derivation of knowledge (14).

Considering this context, the problem addressed by this study is: How to identify and recommend influencers and content published on different social networks and provide data analysis, capable of adding value to decision-making in agribusiness.

1.4 METHODOLOGY

We can apply several techniques to data analysis in OSNs. Among them, those that perform structural analysis of the OSN and those that perform semantic analysis are investigated in this work. Moreover, we can combine structural analysis with techniques related to artificial intelligence to discover patterns not detected. These techniques, together with the semantic analysis of OSN data, can bring new information and trends that would not be detected if these techniques were not combined. In this sense, using machine learning techniques together with semantic models, such as ontologies, can help derive strategic information for the improvement of the dairy market.

[1] https://observatoriodoconsumidor.cnpgl.embrapa.br/

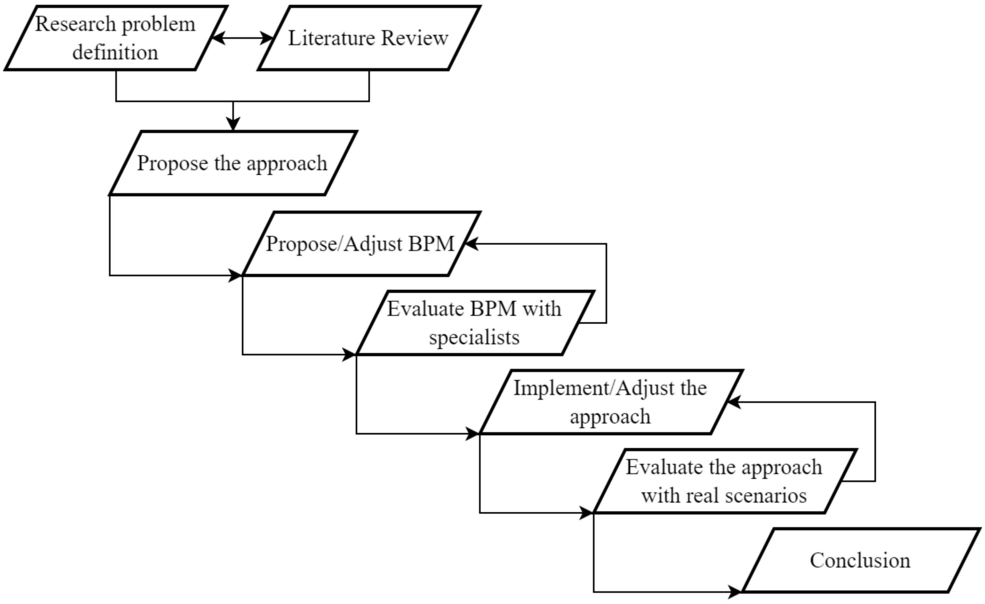
For example, using Machine Learning (ML) techniques (15) we can discover specific users in the OSN, and based on these specific users, using ontologies, discover new relationships between them. Moreover, using these newly detected users in the network, we can identify the most influential user who can publicize the product or spread information to help a company advertise its products (6). We can also detect communities that can be suggested to specific consumers (16). Therefore, ML techniques and semantic models as ontologies (17) may be a path to follow since discovering specific patterns and generating new implicit relationships generates new marketing opportunities.

Therefore, to apply these techniques, we conducted this research (Figura 1) using the Design Science Research (DSR) methodology (18) in two cycles. In DSR, knowledge and understanding of a problem domain and its solution are achieved through the designed artifact's construction and application. Artifact evaluation provides feedback information and a better understanding of the problem to improve product quality and the design process.

The IntelDigitalMarketing, developed as our solution, corresponds to the artifact. In the first cycle, we developed the REDIC architecture (6). The first cycle evaluation conduction generated scientific knowledge. This knowledge helps to evolve and generalize the solution, and we proposed the IntelDigitalMarketing approach in the second cycle.

As a theoretical basis, this conjecture derives from the knowledge acquired in research already carried out, identified by the literature review and the first DSR cycle results.

Figure 1 - Research Steps Conduction


Research problem definition
Literature Review
Propose the approach
Propose/Adjust BPM
Evaluate BPM with specialists
Implement/Adjust the approach
Evaluate the approach with real scenarios
Conclusion

1.5 RESEARCH QUESTION

To verify the feasibility of our proposal, we proposed the following research question (RQ):”**How to foster Agribusiness products by analyzing multiple OSNs?**”. To answer this RQ, we derived two Secondary Research Questions (SRQ)

- SRQ1. What is the marketing effect of the most cited cheese among the communities of the different generations of users who talk about cheese on the OSN? The purpose of this RQ is to answer which cheese is most cited among users of an OSN. Identifying the most talked about the product among communities can help producers enhance the market for similar dairy derivatives.
- SRQ2. Who are the biggest influencers of cheese at OSN? The purpose of this RQ is to highlight in an OSN the most influential users in the dairy segment, specifically in the production of cheeses. Thus, we can take advantage of these users as influencers who can promote dairy derivatives. We can also provide evidence of how producers can use information from influential users as guidelines for developing new products, packaging, and prices, for example.

1.6 OBJECTIVE

The main objective of this work is the specification of a data analysis solution, based on identifying influencers communities in the dairy sector to assist in disseminating information about dairy derivates and presenting guidelines for the marketing practice for producers. The focus is on multiple OSNs, combining Twitter, Instagram and YouTube.

As specific objectives we can mention:

- • Development of an ontology called Artchee-O, with the aim of representing social media related to the dairy products domain.
- • Specification of a process called BPM-IntelDigitalMarketing, systematizing the phases of our solution;
- Development of an architecture called InteldigitalMarketing, able to recommend content and influential users;
- The creation of dashboards that help decision makers to analyze data systematically, supporting strategic decisions.

1.7 DISSERTATION STRUCTURE

The dissertation is organized as follows: chapter 2 reviews the main concepts related to this work. Chapter 3 discusses related works considering social networks and data recommendation analysis related to Agribusiness. Chapter 4 presents the proposed solution. Chapter 5 details a feasibility study considering the dairy derivative domain. Chapter 6 presents the final considerations.

2 THEORETICAL FOUNDATION

In this chapter, we review the main concepts related to the proposal.

2.1 COMPLEX NETWORKS

Complex networks are made up of a set of elements that are related through explicit or implicit associations. These associations can create a complex graph of connections between the elements that make up the network. This complex graph makes networks challenging to understand in some cases, as there may be many connections between elements that are only sometimes consistent (34). This complexity comes from the types of relationships between the elements that compose them.

In computer science, networks and graphs are structurally equivalent. The nodes of a network are equivalent to the vertices of a graph, and the relationships between them are represented by the edges between them (35). Using a complex network not only values the data, but mainly the relationship between them, allowing the derivation of new knowledge from these relationships and specific properties of the network. Thus, data can be represented by nodes of a graph, and we can study their relationships so that, for example, the most relevant node of the graph, or network, can be found.

In many real-life applications, we need to identify highly influential or important nodes in networks with large nodes and complex topological structures. For example, if we want to set up a customer service center, what would be the best location, or if we want to provide free product samples, for whom should we send them? The importance of a node depends on the context of the application. A central node, for example, can help spread information locally, but it will only be as effective when not connected to other influential nodes. Researchers have studied various phenomena and determined which nodes are most important based on application requirements

(3). Several centrality measures were defined to identify these nodes: *degree centrality*[2] , *betweenness centrality*[3], and *closeness centrality*[4], among others.

It is important to highlight the relevant role that social networks play in the modern world. Its analysis as a complex network can help study several current phenomena and thus present advances in several research areas, such as Agribusiness, for example.

2.2 SOCIAL NETWORK AND MEDIA ANALYSIS

Considering the research context of this dissertation, an analysis of Online Social Networks investigating their properties in a complex network context is relevant. Thus, given the volume of publications in an OSN, a major challenge is to find out what information is relevant to a given domain and how to analyze it (19).

OSNs often contain a lot of content and linking data that can be used for analysis. This data can be subdivided into structured and unstructured data, respectively. OSN structured data is usually graph structured. They are modeled based on a social network represented as a graph G = (V, E) where V is a set of vertices or entities (e.g. organizations, people, and products) and E is a set of edges or relationships that connect the vertices through interactions. Structured data is measured through SNA, where a graph analysis application extracts intelligence from such data.

SNAs rely on some techniques such as community detection (63), which is a data clustering problem that groups nodes of a network into communities in a coherent way. One way to quantify coherence is with modularity, which is a measure of the structure of the network. It was designed to identify the density or strength of connection between nodes in a network. Networks with high modularity have dense connections between different modules, but sparse connections between nodes within

[2] It's an measure that individually quantifies how connected a node is in the network, using the number of direct connections of that node (68)

[3] This measure quantifies the degree to which a node is on the shortest path between two other nodes and is capable of acting as a bridge (68)

[4] Through this measure it is possible to quantify the shortest path from one node to another node through the network (68)

individual modules.. We used modularity to examine resulted communities. Communities play an important role in both exploring a network and predicting connections that have yet to be observed.

Another technique applied in OSNs is influence analysis (64), which focuses on identifying experienced users, using measures such as centrality. These experienced users can be opinion leaders because their opinion posts are the ones that can spread considerably faster through the network. For that, we use degree, betweenness and closeness centrality (more details in Section 4.3.2.1).

Link prediction is another technique used to study connections between nodes and predict new relationships between entities.

Unstructured data, on the other hand, is content data shared on OSN, also known as User Generated Content (UGC). This data is usually analysed through Social Media Analysis (SMA) techniques. SMA can use several techniques such as sentiment analysis which uses the textual content of a specific entity to determine the sentiment of the text. Another technique is topic detection, which is typically used to detect emerging topics in a textual dataset to better understand societal concerns and detect trends. This technique is commonly used in social networks together with unsupervised machine learning. Other non-text media can also be used to extract intelligence from content shared on the network, such as audio, video, or image analysis. Image analysis can use the facial recognition technique, which allows for extracting sentimental and demographic data from profile images. The audio analysis uses continuous speech recognition vocabulary or a phonetics-based approach to extract information from unstructured audio data. Video content analysis involves extracting meaningful information from video streams.

SNA and SMA approaches have a vast collection of analysis tools (20, 21). It is essential that new relationships or knowledge can be found through these analysis methods. This new knowledge can then be used to deal with issues such as information filtering and information overload.

2.3 SEMANTIC WEB

"*The power of the Web is in its universality. Access for all, regardless of disability, is essential.*" Tim Berners-Lee (28). In 2001, Bernes-Lee, Hendler and Lassila published an article proposing the use of the Semantic Web as an extension of the current Web. This proposal would involve using technologies to help people perform basic tasks, where data on the Web would be related, allowing computers to perform more “intelligent” tasks, such as understanding the context and meaning of information on the Web (29).

Although the existing Web allows the use of specific techniques to find information, such as keywords, is also needed intelligent applications to understand the meaning of these keywords. Thus, techniques related to the semantic analysis of information began to be investigated, deriving the concept of Semantic Web. Among these techniques, it is worth mentioning the use of folksonomies, in basic contexts where semantic processing is restricted, and ontologies.

A Folksonomy can be considered as a set of keywords that annotate and describe online content. They help to understand and annotate specific concepts on the web, helping in the linkage of web data. However, folksonomies do not use specific rules for deriving new knowledge, as can be done in an ontology (72).

On the other hand, an ontology can be considered a system of categories representing a way of seeing a world and a set of rules that help to relate these categories and find new relationships. Ontologies are widely used to describe a domain consisting of specific vocabularies that generally describe a given reality (31). Its main elements are entities, which describe concepts related to an application domain, properties that describe these entities and associations that link them. In addition, an ontology has constraints and rules that apply to entities, associations, or the model. When processed by inference algorithms, these restrictions and rules allow the derivation of new knowledge associated with the other ontology elements.

Thus, an ontology can represent a reality, providing a taxonomy to interpret and infer new knowledge about a specific data domain (32). We can use this process to support the reasoning process of semantic analysis of data, such as social data from OSNs.

An ontology in the context of the Semantic Web is implemented in an ontological language, such as OWL (Ontology Web Language) (69) and can be queried using the SPARQL language (33). The SWRL (71) language can implement logical rules related to the ontology. In OWL, entities are implemented as classes, associations as object properties and properties as data properties. Other specific constraints can also be detailed in OWL, such as the transitivity of associations or symmetry. More complex restrictions and rules can be implemented in SWRL and queries can be specified in SPARQL.

2.4 FINAL REMARKS

This chapter discussed the main concepts related to the proposal solution, emphazing the state of the art Technologies. We detail the main concepts related to the development of the approach, emphasizing the importance of this research topic. In the next chapters, we present an approach that supports decision making in agribusiness using process management, ontology, SNA and SMA. In addition, we present a systematic mapping carried out to investigate the main works related to the proposal of this dissertation.

3 SYSTEMATIC LITERATURE MAPPING

In this chapter, we investigate the main works related to the proposal of this dissertation. To do this, we conducted a Systematic Literature Mapping (4) to find out relevant works. We analyse the works and discuss research directions.

According to (36), Systematic Literature Mapping (SLM) is an overview of primary studies on a specific topic that aims to identify subtopics that need more primary studies. This study uses a protocol following methodological steps to make the results more reliable. This mapping, therefore, focuses on research associated with the use of crowdsourcing in activities related to agribusiness. The main element of analysis uses data from online social media, in other words, “websites and applications that allow users to create and share content or participate in social networks” (37).

The research problem addressed in this SLM is the need to support agribusiness in consumer market analysis to improve the sector. To tackle this problem, we investigate new strategies to improve marketing in agribusiness.

3.1 **SEARCHING STRATEGY**

In the SLM method adopted in this work (39), the authors propose four strategies that combine database searches in digital libraries with backward snowballing (BS) and forward snowballing (FS) iterative (BS*FS), parallel (BS||FS), or sequential BS+ FS and FS+BS. The authors also conduct a comparative evaluation of traditional digital libraries to find the database with the most significant performance results. They considered the Scopus database the most consistent digital library in terms of accuracy. In addition, the library integrates other digital libraries into its search method, increasing the search reach. However, it is necessary to complement the library with the snowballing process.

In our work, the hybrid strategy adopted is Scopus + BS||FS. In this strategy, an initial set of papers is obtained through Scopus. Then BS and FS are performed in parallel on the same initial set. In other words, articles obtained by BS are not subject to FS and vice versa. We introduced this strategy to increase accuracy without compromising recall (40).

3.2 **RESEARCH QUESTIONS**

The following Systematic Literature Mapping Research Question (SLMRQ) was formulated to conduct the study: “**How can the use of SNA and/or SMA support marketing improvement in agribusiness?**”. This SLMRQ aimed to investigate the techniques used and where these solutions are applied to understand how the solutions can help in decision-making. This SLMRQ was associated with four secondary questions, as shown in Table 1.

From the research questions, we used the PICOC method (41) to define the scope of work and the terms used in the search string, as illustrated in Table 2. The search string, the main terms of the SLMRQs, synonyms, and acronyms were specified and validated with the assistance of an agribusiness expert and control articles (13,42–44) . The identified terms form the following search string: *((("online social network" OR "OSN" OR "social network" OR "social media") AND ("analysis" OR "approach" OR "architecture" OR "analytics")) OR "SNA" OR "SMA") AND ("agriculture" OR "agri-food" OR "agronomy" OR "agribusiness" OR "agro-industry" OR "agricultural business" OR "dairy products")*. Two external researchers revised the protocol. Furthermore, we only cover the SNA/SMA studies applied to agribusiness. From the string, we obtained 234 publications in the period 2017 – 10/2022, after the application of the inclusion filters. The search filters used are detailed in Section 3.3.

Table 1 – Secondary research questions.

ID	Research Question	Goal
SLMRQ1	What are social networks used for data collection in the SNA/SMA agribusiness research community?	Find which social networks are relevant data sources for the SNA/SMA survey in agribusiness.
SLMRQ2	What are the SNA/SMA analysis techniques used in agribusiness studies?	Identify SNA/SMA techniques commonly used for analyzes performed in agribusiness (e.g., sentiment analysis, influence analysis, textual analysis, among others).
SLMRQ3	What are the evaluation metrics?	Highlight which evaluation metrics were used to verify the proposed solution (e.g., accuracy, F1-score, number of publications, among others). This research question aimed to assess the metrics used to evaluate the experiments conducted in the proposed works, i.e., if the work conducted an experiment to evaluate the proposal, which metric was used in this evaluation?
SLMRQ4	Which subdomains of agribusiness are studied?	Find where researchers apply their approaches in the context of agribusiness (e.g., milk and dairy products, food security, and urban agriculture, among others).

Table 2 – PICOC

PICOC	Description
Population (P)	SNA and SMA solutions
Intervention (I):	Agriculture, Agribusiness
Comparison (C):	Not applicable
Outcome (O):	Approach
Context (C):	Crowdsourcing, OSN

3.3 **INCLUSION AND EXCLUSION CRITERIA**

Due to the large number of documents retrieved from the digital library (Scopus), we used the inclusion and exclusion criteria to select only potentially relevant articles returned from Scopus. Then, we applied the criteria to eliminate papers not related to the goals of this mapping. We used the Parsifal[5] tool to support the mapping execution. The inclusion and exclusion criteria used in this work are shown in Table 3.

Table 3 – Inclusion and Exclusion Criteria

Inclusion Criteria	Description
1	Search by title and abstract.
2	Studies published between 2017 - 10/2022.
3	Studies we have full access to using the university (UFJF) credentials.
4	The article was published in a Journal or Conference.
5	Papers are written in English.
Exclusion Criteria	**Description**
1	SNA and SMA studies do not have OSN as a data source.
2	Studies that are not applied to the agribusiness sector.
3	Books, sections, and book chapters, considering that they are not peer-reviewed.
4	Review, papers.

We selected studies at two levels: (i) a new search was performed, adding the inclusion criteria as filters in the advanced search by Scopus; and (ii) a reviewer guided by the exclusion criteria performed a selection considering the title and abstract. It was verified whether both explicitly reference social media or social networking services in Agribusiness. Records at this level were kept when there was doubt about their

[5] https://parsif.al/

relevance – just reading the title and abstract was insufficient for the evaluation. As a result, additional sections of the articles were read.

3.4 **QUALITY ASSESSMENT**

According to the guidelines proposed in (36), researchers can develop a quality assessment for primary studies. The evaluation serves as a guide to understanding the results. Each article selected after the inclusion and exclusion criteria was evaluated considering the quality Assessment Questions (AQ). These questions, presented in Table 5, were developed based on (45,46). For each question, we assigned the value 1 if the answer was "yes", no value was assigned if "no" was answered, and 0.5 value if the answer was "partially".

We established a predefined questionnaire (Table 4 presents details) and the AQ questions (Table 5) and highlighted information such as the social media source used and the evaluation metrics to answer the questions. The publications were categorized and tabulated according to the questions, extracting their information related to the questionnaire. This technique helped us to detect and validate the data extraction results and settle any discrepancies. The questionnaire can be accessed in [6].

[6] https://forms.gle/Xb56pN8ybcSPYXbu7

Table 4 – Questionnaire details.

Questions	Example answer
What is the name of the paper?	#Eggs: social and online media-derived perceptions of egg-laying hen housing
Does the study use SNA and/or SMA?	SMA
Which SNA or SMA technique is used?	Sentiment Analysis; Age and Gender Estimation; Time Analysis
What is the most used evaluation metric?	Statistical methods (sum, average, percentage, frequency)
What is being analyzed?	Eggs
Which social media site is used?	Netbase Analytics Platform

Table 5 – Quality Assessment Questions.

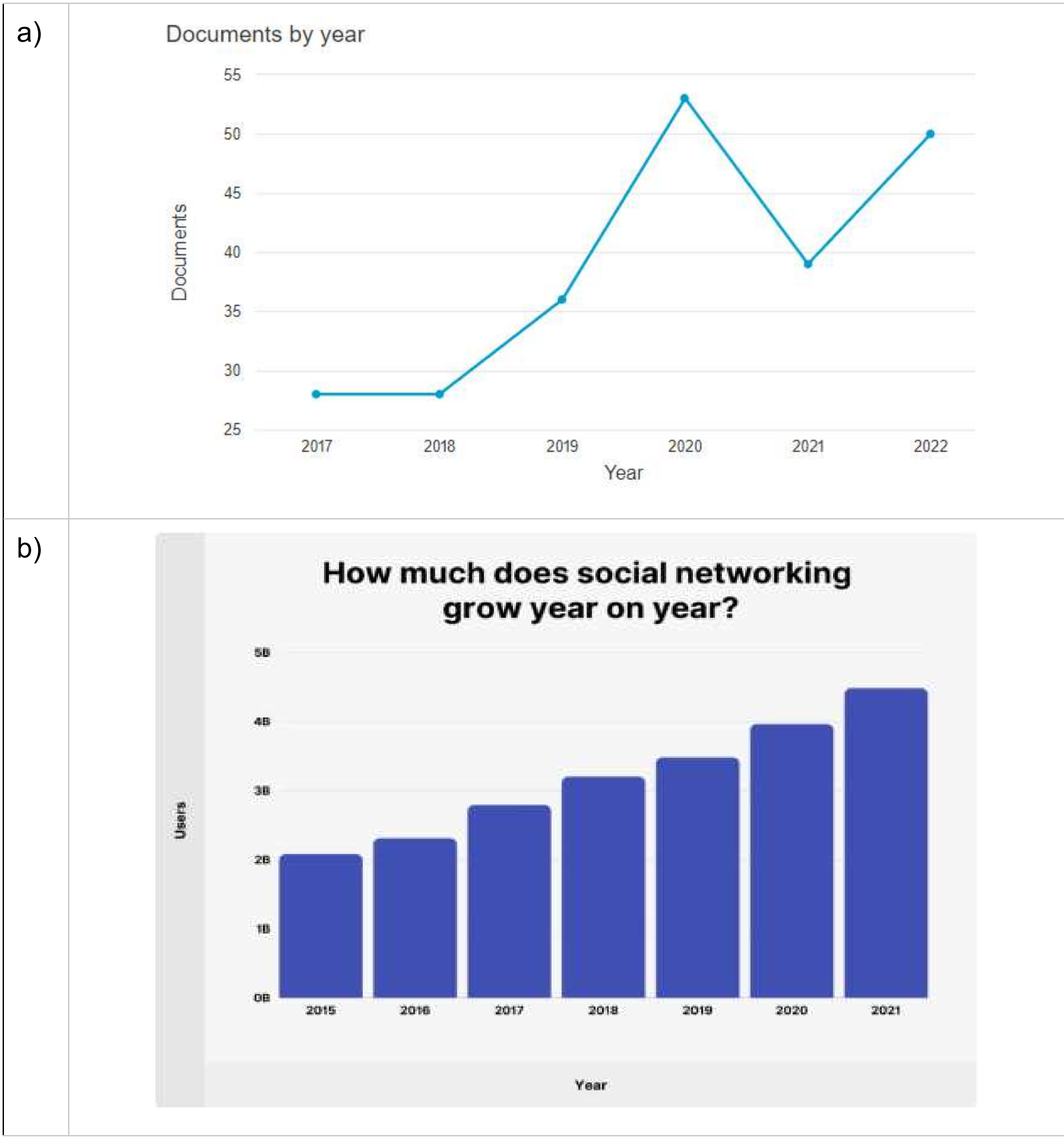
Assessment Questions	Description
AQ1	Is the purpose of the study clear?
AQ2	Are the uses of the techniques justified and clearly described?
AQ3	Are data collection methods adequately described?
AQ4	Is the collected data adequately described?
AQ5	Do the authors discuss limitations, threats to the validity, and reliability of the study results?
AQ6	Are research questions answered clearly?
AQ7	How clear is the interpretation of the data collected and the conclusions in the text?

3.5 **RESULTS**

Based on the inclusion criteria (first level), 234 publications were returned from 2017 until October 2022. Figure 2 (a) shows an increasing number of articles published between 2017 and 2020. This growth can be due to the growing popularity of OSNs over the years, prompting the researcher´s interest. In 2020 there were 3.96 billion active users on OSNs worldwide, and in 2017 there were only 2.79 billion active users

– an overall increase of 41.63% Figure 2 (b). However, concerning the number of articles published, the year 2021/2022 was lower than 2020. This fact could be due to the COVID pandemic that started in 2020. According to (47), the search for studies related to the disease has increased.

Figure 2 – (a) Distribution of publications per year and (b) Global social media growth rates per year[7].

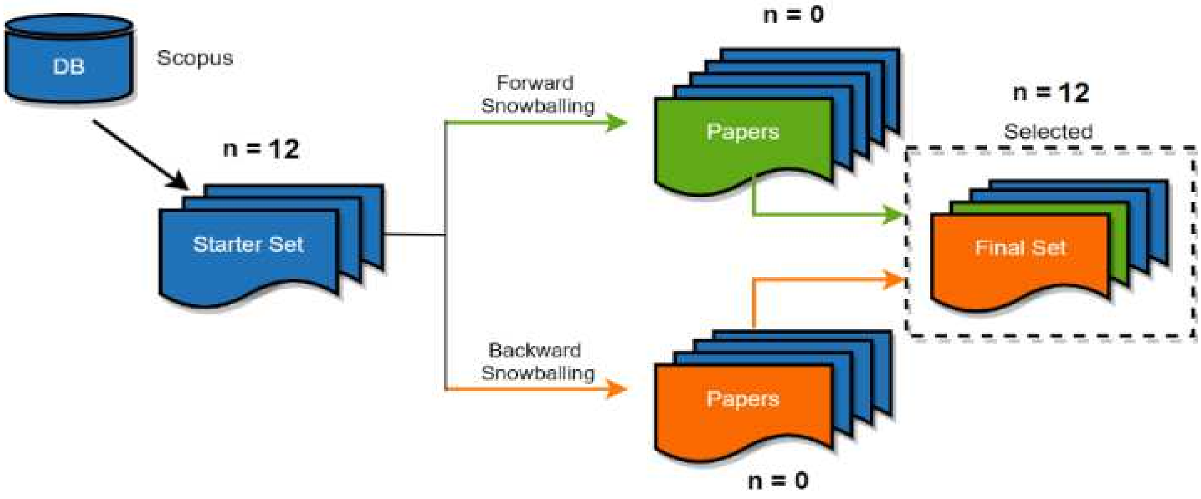


[7] Accessed in 07/21/2021: <https://backlinko.com/social-media-users>

To determine study eligibility, all publications that used the identified social media sources, based on the definition given in (37), were considered, including blogs, news, and other user content-sharing sites (e.g., online forums). At the second level, described in Section 2.5.3, few studies considered OSN a collection source for SNA and SMA in agribusiness. Considering publications that exclusively use SNA or SMA (~23%), ~69% used traditional data collection techniques, such as questionnaires and interviews, using OSN as a means of communication. Others (~2%) did not use social media to collect data. They collect data through direct questionnaires. Therefore, these studies were discarded according to exclusion criterion 1 shown in Table 3. Finally, 12 papers were selected as a set of studies directed to the BS||FS.

A support tool[8] configured to carry out the process using Scopus as a database was used in the snowballing process. The Forward Snowballing (FS) process execution resulted in 13 papers, and the Backward Snowballing (BS) process 408 papers. After applying the inclusion criteria, 12 papers remained in FS set and 30 in BS set. However, no papers were left in FS and BS when using the exclusion criteria. Thus, out of 234 articles, only ~5% (12 papers) remained in the final set. This significant reduction can be explained by the number of false-positive studies captured in Scopus through the word "agriculture" and its variations in the search string. The process can be seen in Figure 3.

Figure 3 – Hybrid methodology structure (Scopus + BS||FS)



[8] https://github.com/JoaoFelipe/snowballing

3.5.1 Data Extraction and Analysis

The papers selected were analyzed and the data extracted to address the research questions.

SLMRQ1. What are social networks used for data collection in the SNA/SMA agribusiness research community? We separated the social media sources for data collection from the questionnaire used in the Quality Assessment to answer this question. This list identified the number of studies (~67%) that use data from a single social media source and the number that use more than one source (~33%). We identified that most researchers prefer to use only one social media source for their research. According to the extracted data, Twitter (~67%) was the most used source, while YouTube, Facebook, WeChat and Sina Weibo were the least used (~8%). The last two OSNs are popular in China and are growing[9.] According to Statista[10] ranking of the most popular OSNs worldwide in January 2022, WeChat and Sina Weibo are among the top ten. Being an OSN used for opinions and considering that short texts are easily mined and processed by its free API, Twitter is the most popular among the research community. Furthermore, it was also identified that some studies (~27%) consider the use of social media analysis platforms such as Netbase[11] (~18%), which has several ONS as a data source (e.g., Twitter, Reddit, blogs, forums, and more), LikeAlyzer[12] (~8%) for Facebook, Twitonomy (~8%) for Twitter and Meltwater[13](~8%), which also has several OSN as a social media source. These analyses can be seen in Figure 4, which illustrates the results obtained for the analysis of SLMRQ1, showing a ranking of the most used OSNs among the community of researchers in agribusiness.

[9] Accessed at 02/04/2022: <https://www.statista.com/statistics/255778/number-of-active-wechat-messenger-accounts/>

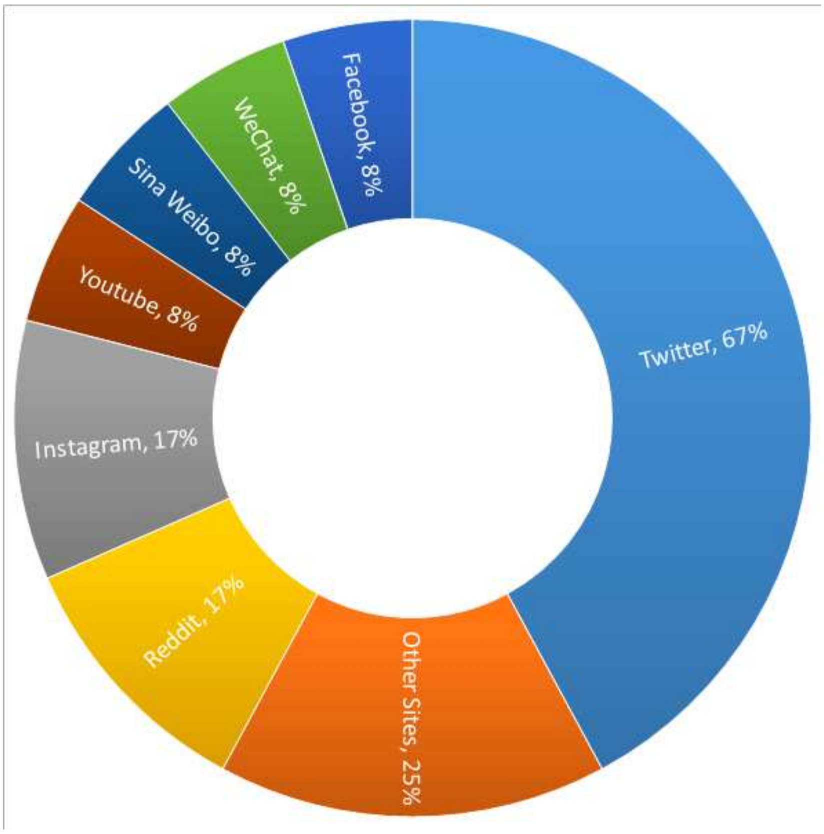
[10] Accessed at 02/04/2022: <https://www.statista.com/statistics/272014/global-social-networks ranked-by-number-of-users>

[11] https://netbasequid.com/

[12] https://likealyzer.com/

[13] https://www.meltwater.com/

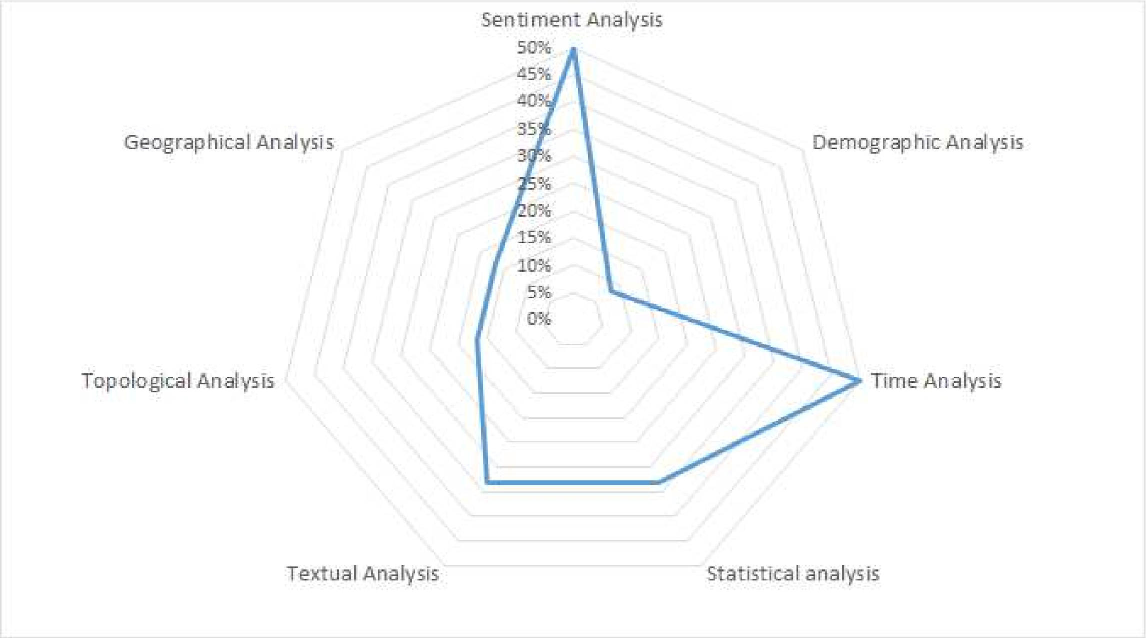
Figure 4 – Popularity of social media sources among researchers.



SLMRQ2. What are the SNA/SMA analysis techniques used in agribusiness studies? We investigated the studies considering which techniques were used to analyze media and social networks to answer this question. We identified that studies using SNA (~25%) were the minority. SNA was used in studies as the sole means of analysis (~8%) or combined with SMA (~17%). In other words, studies that consider SNA tend to consider the use of SMA to solve problems. These studies use textual analysis of the collected data and perform a topological analysis of the most cited words. Topological analysis was the only identified SNA method through the form used to model a keyword network and find trends (48,49). It is possible to observe that the two works that used this type of analysis also used textual analysis. Both works first use textual analysis to find the most frequent keywords. Then, they use topological analysis to understand one keyword's relationships with another. As for the SMA methods, the following were identified: (i) time analysis, where a timeline of social media is made, identifying the number of publications during a period (13,42,50,51); (ii) textual analysis, where social media keywords were studied using natural language processing (NLP) (48,49,51,52); (iii) statistical analysis, where the studies used

hypothesis tests, means and percentages (43,53–55); (iv) sentiment analysis, which aims to identify and extract subjective information from social media by combining NLP and machine learning techniques to assign weighted emotional scores (13,42,50,51,55,56); (v) geographic analysis, where media upload coordinates are used to map their geographic locations (50,51); and finally (vi) demographic analysis, where age and gender of the social media user are estimated using their first and last name as input (13). Figure 5 illustrates this analysis and shows a radar graph of the popularity of the SNA and SMA techniques. In the graph, it is possible to see in the geographic, topological, and demographic analysis that these areas deserve attention. These areas can be better explored as few works have been identified.

Figure 5 – SNA and SMA techniques popularity radar.



SLMRQ3. What are the evaluation metrics used? We analyzed the selected papers to extract the researcher's metrics to understand how the methods proposed in primary studies were evaluated. Most articles (~83%) (13,42,43,49–55) used statistical methods such as total publications, means, and percentages to describe the results and evaluate the proposal. However, none reported an experiment that evaluated the proposed method. Only one of the primary studies had a section describing the proposal evaluation (56). This study used the following metrics: precision, recall, F1-score, and Area Under Curve (AUC). These metrics were used to validate the machine

learning models to classify media data into a sentiment. Finally, only one study in the final set did not use any method to evaluate the proposed approach (48).

SLMRQ4. Which segments of the agribusiness sector are studied? We analyzed the studies to understand the agribusiness subsectors where the SNA and SMA techniques are applied. We found that each work considered a specific segment in agribusiness. In (13), the authors studied the consumers' view on the production system for eggs and laying hens. The authors also exemplify how monitoring OSN can help decision-makers manage agribusiness marketing food systems through the proposed method. In (42), the authors analyzed agricultural markets over 27 months, searching for agricultural market-related keywords. They provide valuable insights about agricultural markets, including the public's view of the livestock industries and the risk of zoonotic diseases. In (43), the authors examine companies' engagement in the olive oil sector in the OSNs and compare organic and non-organic operators. According to the authors, organic food products in Spain face commercial problems due to factors such as the considerable price differential between organic products and their conventional equivalents.

The study reveals statistically significant differences in the engagement and use of OSNs by non-organic and organic operators. In (53), the authors combine data from 13 sites in 11 low-income countries to study how various social capital scales relate to household food security outcomes among smallholders. The authors conclude that social network theory correlates household food security with multiple social capital scales, both within and outside the household. This social capital can be either a link (within groups) or a bridge (between groups) with different implications for how the structure of social capital affects food security. The article (49) presents the first content analysis on the Czech Twitter OSN in the context of agriculture in general. The authors identified a prevalence in tweets about biofuels, the rapeseed plant, and politics. Furthermore, they conclude that robot accounts created a significant proportion of tweets. In (54), the authors share their experience using the WhatsApp platform for communication and data collection to monitor and evaluate the sweet potato value chain. The article (48) drew only on data from public newspaper reports and a sample of the social networks used by urban food networks in Bristol - a city with a well-developed urban agriculture movement - to explore how activists in urban agriculture food use OSNs during 2015. The authors intended to inform debates on urban agriculture and contribute to discussions on its growth. In (52) the authors studied the influence of COVID-19 on China's agricultural economy. In (55), the authors address the problem caused by wild pigs for agriculture and the environment. Through SMA, the authors find evidence of a need for more information on best practices for safety, such as the risk of zoonotic diseases caused by wild pigs. In addition, they describe the importance of understanding the influence of social media on people and opportunities for management agencies, such as messages in public health campaigns. In (50) the authors explore Artificial Intelligence (AI) in agriculture. Based on SMAs, the authors conclude that AI techniques in agriculture are positive. In the work presented in (51), the authors focus on how SMA can support government authorities in predicting damages related to the impacts of natural disasters in urban centers. The study uses SMA in Twitter crowdsourced data using the keywords “Disaster” and “Damages”. The study's methodological approach employs the social media analysis method and performs sentiment analysis and textual content of Twitter messages.

Finally, the work of Salim et. al. (56) uses public opinion in OSN to determine whether Smart Agriculture or Agriculture 4.0 is implemented in Indonesia. Therefore in answer to SLMRQ4, ten segments were identified: eggs and laying hens, agricultural markets, olive oil, family food security, agriculture in general, sweet potato, urban agriculture, agricultural economy, intelligent agriculture, wild pigs, impacts of natural disasters in urban centers and AI in agriculture. Several agricultural segments have not been explored in the OSN context, such as milk, its derivatives, and other commodities.

3.5.2 Threats to Validity

This systematic mapping aimed to provide an overview of the literature regarding using OSN, identifying, categorizing, and analyzing SNA and SMA solutions in the agricultural domain. However, some threats to validity and limitations can influence the results.

The search string as a threat to the construction validity should be mentioned (see Section 3.2). We defined the main terms of the SLMRQ, synonyms, and acronyms considered adequate to make the string as comprehensive as possible. However, some terms may have yet to be considered in the string. To address this, several tests were carried out with the terms, and as we carried out the searches, versions were generated to decide on the best string. In addition, experts reviewed them, and control papers were used. As a result, this threat was mitigated.

All formulated conclusions, and results found in this SLM, have traceability. However, biased data extraction from selected articles may threaten the conclusion's validity. In other words, we may have included papers in the final selection set that can be a false positive. We used a predefined form of data extraction to mitigate this threat.

We only included articles that could be accessed using our university (UFJF) credentials. This restriction can also be considered a threat to validity. However, with the snowballing technique, this threat has been mitigated.

Regarding threats to internal validity, the SLM selection process (see Section 3.1) was conducted by only one researcher, and it may be challenging to include all relevant publications in the research. Furthermore, it is challenging to ensure that all topic-related concepts and relevant articles have been included in this study, despite the care and effort taken. However, the snowballing technique was used, which helped include pertinent new works.

3.6 **RESEARCH DIRECTIONS**

As social media and network analytics evolve in agribusiness, it is possible to identify the various social media sites and analytics platforms used. The most popular, if not the most representative, social media site is OSN Twitter.

We observed that social media data are primarily collected using Twitter as a source, which is used in ~64% of the selected studies. Thus, the Twitter platform is the most used by researchers. In general, data from OSNs are the most used, either as a single source of collection, in conjunction with another OSN, or through analysis platforms.

The most used analysis technique in agriculture is the machine learning technique in sentiment analysis, which is used in ~50% of the selected studies. However, we find increasing activity in both time-based techniques and statistics. We also observed that precision, recall, and accuracy score are the evaluation metrics most used by the selected studies to evaluate their experiments.

After analyzing the selected studies, we identified some research gaps that are important to be investigated. The data collected from the Twitter platform is the most used dataset in SNA and SMA applied to agribusiness compared to other social media platforms. This preference for Twitter is mainly due to the ease of access. However, it is necessary to advance the research with the growth in the importance of other social media and mechanisms for extracting information concerning different media types, such as photos and videos. This deficiency limits the generalization of the results obtained through the analyzes. Therefore, analyzing other important OSNs such as Instagram, YouTube and TikTok could bring important insights for agribusiness marketing.

Thus, research directions should focus on using multiple OSNs combined with the application of information extraction and analysis techniques in the various segments of agriculture, especially the underexplored segments, for example, the milk and dairy segments. We can also explore techniques such as topological analysis and semantics, which can assist in extracting information from new media such as augmented reality.

Integrating information from various OSNs, analyzing this data, correlating findings, and directing marketing strategies is an important research direction. Knowledge discovery from this correlation of information can help discover implicit relationships between data. In addition, traceability techniques to verify the integrity of information is also important.

Considering the SLMRQ: “How can the use of SNA and/or SMA support marketing improvement in agribusiness?”, the results of SLM can provide directions to follow. The investigation using multiple OSNs is a path to follow. Machine Learning techniques are also identified as a technique to be used. All these research directions are important and should be investigated in future research.

As a result of our investigation, we consider that the following techniques should be investigated: i) the use of multiple OSNs, combining their results in some way, ii) the use of intelligent data analysis techniques, emphasizing ML techniques, but also investigating other techniques such as semantic and structural analysis, given the structure of an OSN, iii) adequate visualization mechanisms to leverage the commercialization of specific products.

Analyzing the interaction between users in multiple OSNs contributes to identifying consumers with similar characteristics in different OSNs, using SNA techniques. It is also possible to investigate content trends for each community of each OSN, enabling the search for the most talked about information about products, using SMA techniques. As a result, with the combination of SMA and SNA techniques, agribusiness producers and researchers can identify specific product demands, and the consumers that must be reached.

3.7 FINAL REMARKS

We mapped the state of art in using SNA and SMA to leverage agribusiness. The study identified online social media sources, platforms, and techniques for analyzing information, assessment metrics, and the agricultural segments covered by the primary studies.

The SLM showed that more primary studies on SNA and SMA in agribusiness are needed. For example, in the SNA context, few works use community analysis and influence analysis techniques. In addition, one of the untreated agricultural segments, which is of significant economic importance, is milk and dairy products. There is a need to assist marketing in the agricultural sector using analytical methods and opportunistic crowdsourcing.

The results obtained by the primary studies selected in this SLM, together with insights into online social media data analysis, reinforce how social media monitoring can complement traditional methods to inform agricultural producers and consumers about marketing opportunities and regulation of agribusiness. In the next chapter we detail an architecture that can deal with these issues.

4 INTELDIGITALMARKETING SOLUTION

Considering the results of SLM, we proposed the IntelDigitalMarketing solution. For its development, functional requirements (FR) were proposed, considering the research directions pointed out in the SLM: (i) FR1– Automatic discovery of new knowledge to support marketing strategies; (ii) FR2 – Focus on the relationships and connections between OSN users through social media; (iii) FR3 – Recommend trends and new relationships between users of digital influencer communities; and (iv) FR4 – Support the sharing of knowledge generated with other applications (e.g., decision-making agents, APIs). As non-functional requirements (NFR): NFR1 – Scalability (to support other social networks): The solution must be scalable as the volume of information on OSNs can expand; NFR2 – Interoperability: The solution must be able to interoperate with other applications and platforms. Moreover, the solution can be used in various application domains.

Section 4.1 details the steps to develop the solution according to the Design Science Research methodology. In subsection 4.2 we detail the activities related to IntelDigitalMarketing use through the BPM-IntelDigitalMarketing, encompassing the main processes and how activities are performed through the process flow. In subsection 4.3 we detail the IntelDigitalMarketing architecture and its modules.

4.1 DESIGN SCIENCE RESEARCH STEPS

In the Design Science Research (DSR) methodology, artifact evaluation provides feedback information and a better understanding of the problem to improve product quality and the design process. In this vein, the IntelDigitalMarketing, developed as our solution, corresponds to the artifact.

Each DSR cycle generated scientific knowledge. This knowledge helped to build new versions of our solution. The REDIC architecture (6) was developed as our solution in the first DSR cycle. In this first cycle, the evaluation conduction generated scientific knowledge. This knowledge helped to evolve and generalize the solution, and we proposed the IntelDigitalMarketing solution in the second cycle, considering multiple OSNs and new intelligent techniques, combining ML, ontologies, and complex network analysis. Therefore, the research contribution is the proposal of a solution that identifies and recommends influencers, published content, generates new relationships between influencers and content to collaborate with the marketing strategies. To verify the solution's feasibility, we used the dairy derivate market to provide strategic information that can foster marketing strategies in this domain. As a theoretical basis, this conjecture derives from the knowledge acquired in research already carried out, identified by the literature review (Chapter 3) and by the first DSR cycle results.

Table 6 details the methodology steps to drive the IntelDigitalMarketing development, considering the DSR guidelines.

Table 6 – Design Science Guidelines adapted from (18)

Guidelines	Approach
Development of the approach as an artifact	We proposed IntelDigitalMarketing to identify and recommend influencers and published content related to a specific domain to collaborate with the marketing strategies, using ML techniques, ontologies, and complex network analysis. IntelDigitalMarketing aims to implement strategies for detecting published content and influential user communities on social networks for recommending new content and influencers in a specific domain market.
Problem Relevance	We conducted a hybrid SLM and identified that the literature discusses the necessity of more accurate and representative directions for agribusiness marketing, which require less time and allow a forecast of market trends, such as SNA and SMA techniques. However, the literature do not discuss using ML techniques with ontologies and complex network analysis to support marketing. At the same time, the works demonstrate a real agribusiness need considering marketing innovation.
Approach Evaluation	To assess the feasibility of our solution, we carried out a historical research study in a real context. In the first DSR cycle, we conducted historical research with the Twitter data (6). Based on this first cycle, we identified the value of combining data from multiple OSNs, to improve the results. Therefore, in the second cycle, we conducted historical research using the Twitter, Instagram and YouTube (multiple OSNs).
Research Contributions	As a contribution, we developed an innovative solution based on identifying influencers' communities to assist in disseminating information. This solution was customized to the dairy derivatives, presenting guidelines to foster products dissemination.

Guidelines	Approach
Research Rigor	This work details the steps used to develop the solution, according to the DSR and evaluation in two cycles (REDIC and IntelDigitalMarketing) using a historical research approach.
Search Process	To identify the research innovations in the area, we conducted a literature review, exploring how SNA and/or SMA can support agribusiness. Furthermore, we proposed a first solution (REDIC). We continuously improve the solution (IntelDigitalMarketing) through the historical research studies conducted. As a result, scientific knowledge was constructed.
Research communication	The ontological model and the solution prototype can be found at https://github.com/nedsons/.

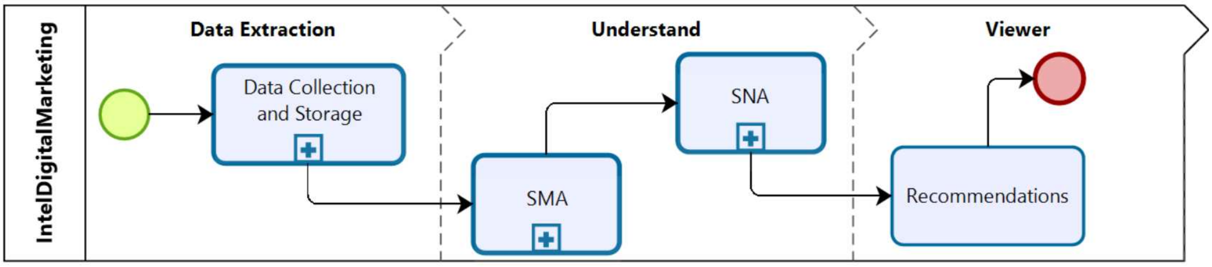
4.2 BPM – INTELDIGITALMARKETING

As stated before, to support information dissemination, the IntelDigitalMarketing solution encompasses activities and sub-processes to detect published content and communities of influential users on social networks and then recommend new content and influencers in a specific domain. To better explain these activities and sub-processes, we use Business Process Management (BPM) (57), as a mechanism to specify the activities flow, considering that BPM is an adaptive management approach designed to systematize and facilitate organizational processes.

With BPM, we specified the BPM-IntelDigitalMarketing process. BPM-IntelDigitalMarketing presents the activities and processes that must be executed considering any domain that the approach can be applied. The aim is to explain its use for any domain that needs digital marketing recommendation. The main process can be seen in Figure 6. We used the Bizagi Modeler[14] tool to model the activities.

[14] https://www.bizagi.com/pt/plataforma/modeler

Figure 6 – BPM IntellDigitalMarketing main process



The workflow in Figure 6 shows the main phases and sub-processes of the BPM-IntelDigitalMarketing. The Data Extraction phase retrieves data from OSNs. The Understand phase selects relevant data to remove noise and poor-quality data, separate it into relationships and content, and employs data analysis methods to classify the relevant data. This phase also creates new relationships between data, generating new knowledge. The Viewer phase recommends the findings of the second stage in a meaningful and targeted way. In the following sections, we detail each phase.

4.2.1 **Data Extraction Phase**

The **Data Extraction** phase involves obtaining relevant social media data by extracting data from OSNs and pre-processing the data using ontology and inference algorithms[15] to select the relevant information. IntelDigitalMarketing currently uses Twitter, Instagram, Facebook, and YouTube OSNs at this stage. Figure 7 details the data collection and storage subprocess, which uses a folksonomy[16] related to the domain where the content recommendation will be performed. This folksonomy is mapped to an ontology that assists in content search, based on the semantic links specified in the ontology. This sub-process also has two other subprocesses: **(i)** creation and instantiation of the folksonomy in the ontology; and **(ii)** data extraction using the ontology.

Subprocess **(i)** starts with the verification of the existence of an ontology (instantiated with the folksonomy concepts) for the target domain (Figure 7-A). If it does

[15] These technologies will be explained in the following sections.

[16] as detailed in chapter 2, folksonomy is a user-generated system of classifying and organizing online content into different categories by the use of metadata such as electronic tags.

not exist, the subprocess to create a folksonomy is started (Figure 7-B). The folksonomy creation sub-process is optional. It aimed to specify a folksonomy, which is a set of keywords organized into categories that describe a specific domain. This folksonomy is created when there is no pre-existing ontology for the domain. The creation of the folksonomy is based on keywords mined on sites related to the domain with the help of domain experts (Figure 8-A and 8-B).

The folksonomy is instantiated in the ontology with the ontology's population subprocess (Figure 7-C). The sub-process **(ii)**, considering the folksonomy instantiated in the ontology, obtains information from the OSNs through a sub-process called Wrapper (Figure 7-E) that searches for keywords in the folksonomy (Figure 7-D). The Wrapper subprocess uses an API for data extraction on OSNs. The collection subprocess ends with the instantiating the mined publications into the ontology. Figure 9 shows an example of the "Coalho" group and its keywords instantiated in an ontology for the dairy derivative domain. . In total, there are 211 individuals instantiated in the ontology, with 19 cheese categories and 180 keywords.

If there are a previous ontology in the domain, it can be used. We need only to extend and/or add specific semantic rules to mine data in OSNs. In Chapter 5, examples of rules for the domain of dairy derivative products will be presented. With the ontology specified and validated, it is instantiated with the mined publications of the OSNs (Figure 10) and with the information generated during the Understand phase, detailed next.

Figure 7 – Data collection and storage subprocess

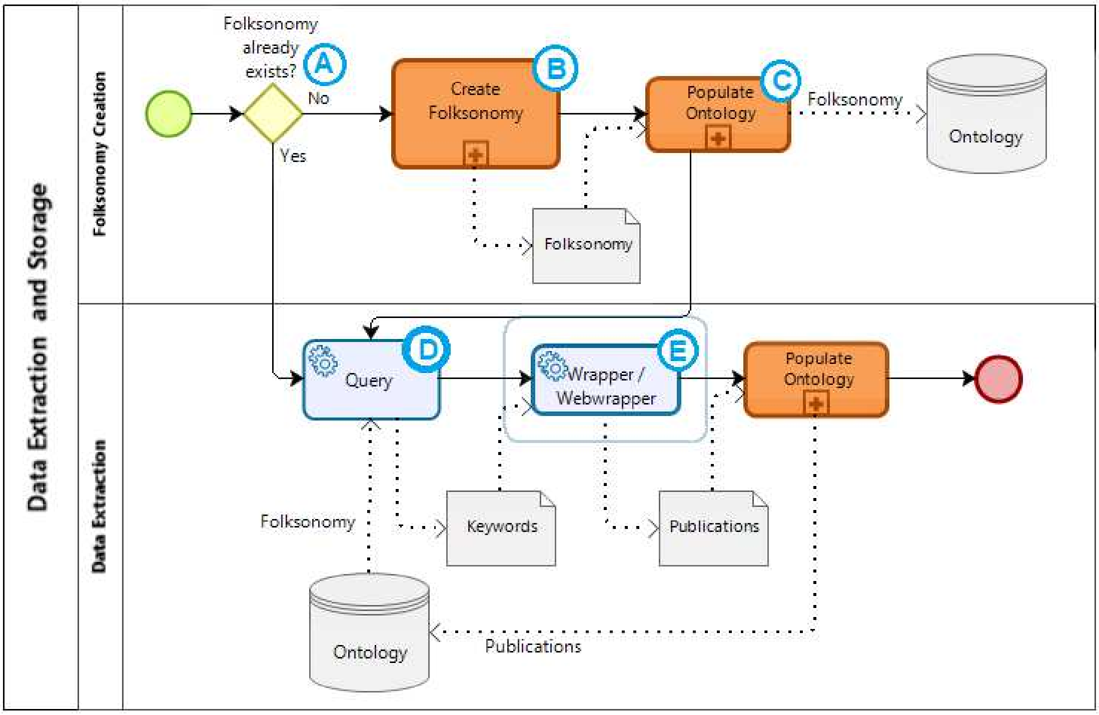


Figure 8 – folksonomy creation subprocess

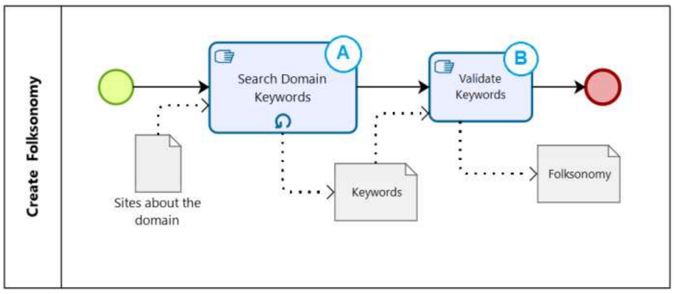


Figure 9 – Coalho group and its keywords in the ontology

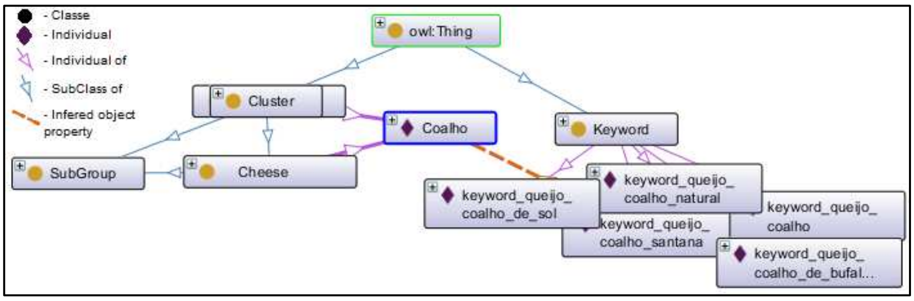
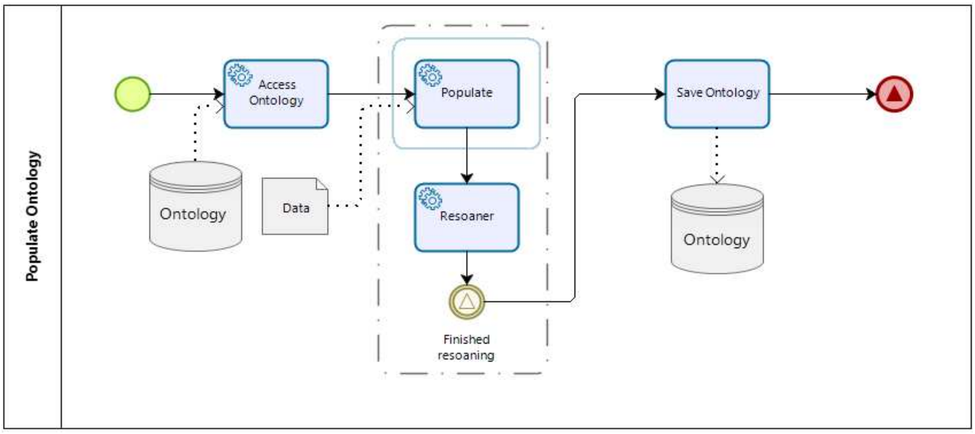

- Classe
- Individual
- Individual of
- SubClass of
- Infered object property
owl:Thing
Cluster
Coalho
Keyword
SubGroup
Cheese
keyword_queijo_coalho_natural
keyword_queijo_coalho_de_sol
keyword_queijo_coalho
keyword_queijo_coalho_santana
keyword_queijo_coalho_de_bufal...

Figure 10 – Populate ontology sub process.



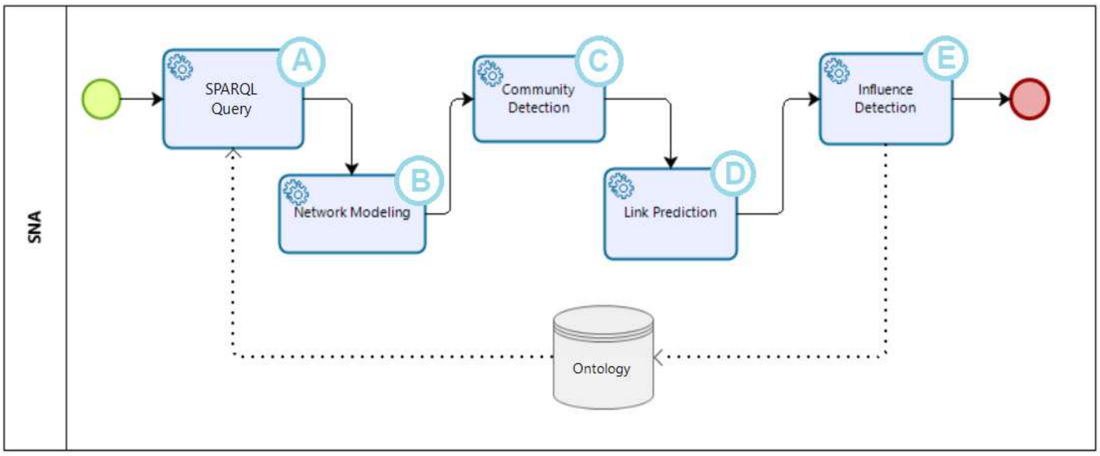
4.2.2 **Understand Phase**

The Understand phase encompasses the processes for structural analysis of the generated network, involving SNA and SMA techniques.

4.2.2.1 SNA

Using the mined publications, the user network OSN is modeled (Figure 11-A and 11-B) from the ontology's data. This sub-process detects user communities that are interactively closer and predicts links between them considering their communities. In addition, it quantifies the influence of these users. The structure of the social network is created and analyzed.

Figure 11 – SNA Subprocess.

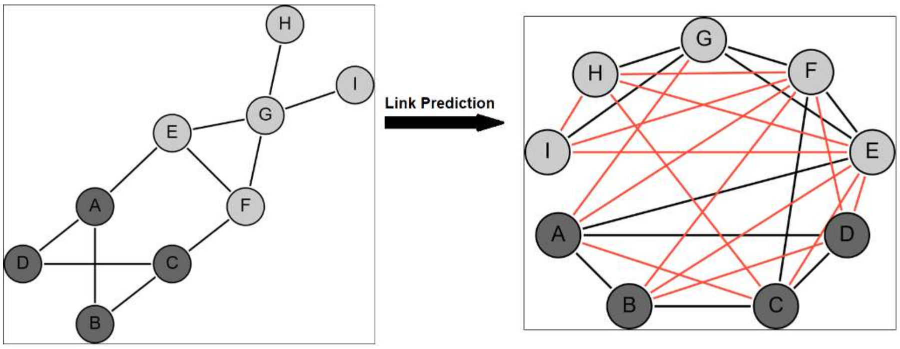


The result of the social network modeling activity (Figure 11-B) is a graph G (V, E) where: (i) a node vi ∊ V indicates a user; (ii) an edge e_{ij} ∊ E indicates an interaction between node 'v_i' and 'v_j'. Suppose that an interaction between users 'v_i' and 'v_j' is found in the instantiated publications. In this case, an undirected 'e_{ij}' edge will be created to connect the two nodes, i.e., if a user mentions another user. We choose to create an undirected edge considering that we are interested in describing the interaction between two users without focusing on their orientations. In our model, each edge has a weight that considers how much users liked the post. Thus, community detection (Figure 11-C) is executed, grouping the generated network into potentially overlapping communities with highly interconnected nodes. Networks with a high modularity score will have more interactions within a community but less interaction with other communities. This type of interaction helps to obtain closely connected components.

Link prediction (Figure 11-D) uses communities to predict edges between nodes, whether within the same community or not. The edge prediction technique used is called Community Common Neighbor (58). We use this technique because it assumes that the network has the concept of communities. For example, groups of people that work in the same area. The technique emphasizes that those nodes belonging to the same community have considerable chances of interconnecting. This function calculates the number of neighbors in common for each pair of nodes. A bonus is added for each common neighbor belonging to the same community as the analyzed pair of nodes.

Figure 12 presents an example of link prediction using the Soundarajan and Hopcroft method (59). The nodes have been separated into two colors representing the communities, which are found by the community detection function. The black lines in Figure 12 are existing relationships. The red lines are possible relationships calculated by the link prediction function, where each line has a score indicating the strength of the relationship.

Figure 12 – Link prediction.

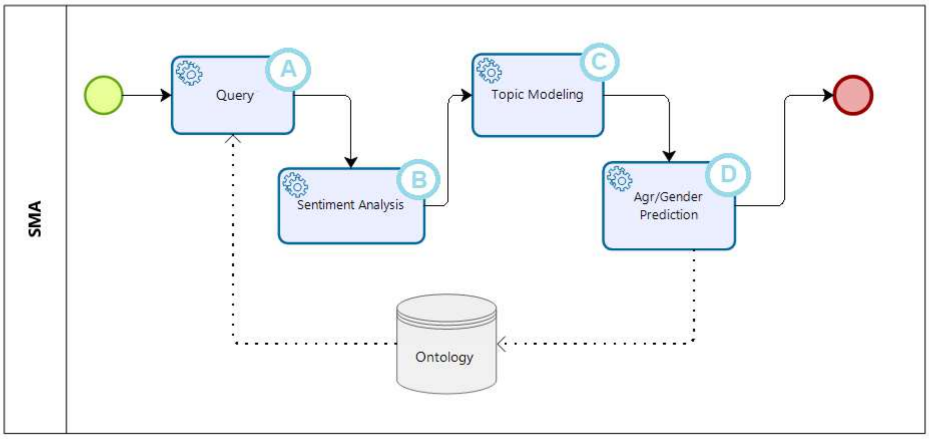


Influence detection (Figure 11-E) detects user influence using network metrics. A SubGraph (SG) is extracted based on a set of nodes of the graph G. In IntelDigitalMarketing, SG corresponds to the most populous community in the network. With the generated SG, an analysis of the centrality of the network is carried out, to find influencers that are the nodes with the highest DC, CC and BC scores. After this activity, the sub-process instantiates all the information generated by the SNA sub-process into the ontology.

4.2.2.2 SMA

The SMA subprocess classifies the instantiated content in the ontology. The ontology is accessed (Figure 13-A) to search for social media content, i.e., text and images. If necessary, we can use filters to return only publications related to one category. After that, the sentiment analysis sub-activity is performed (Figure 13-B), where the textual content is classified between positive, negative, and neutral feelings. The aim is to identify sentiments in the collected media text and classify them based on their polarity to drive product-related marketing strategies.

Figure 13 – SMA sub process.



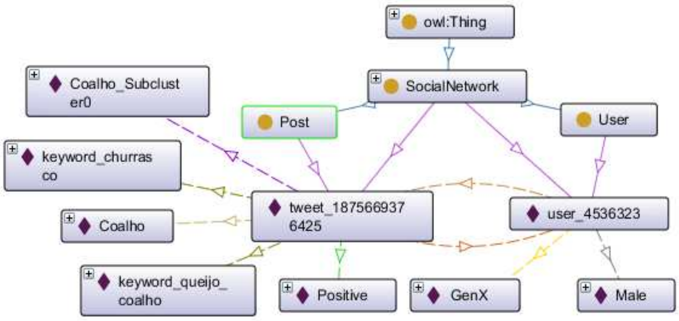
In the following, we carry out the topic modeling activity (Figure 13-C), which analyze the collected publications text, by assigning themes, and organizing the publications into groups based on the assigned topics. Themes are used to provide consistent labels and can be used to support marketing in discovering implicit knowledge. The activity automatically generates topics based on similar posts.

An ML algorithm is used in this step to identify the optimal number of topics and document, and the main weighted terms that contribute to each topic. This activity generates a list of topics, their respective keywords, and the ML model implemented by the algorithm. Then the ontology is instantiated with each topic and keyword. Posts can contain more than one topic, but there is always a single predominant topic.

Finally, in the age and gender prediction activity (Figure 13-D) users' profile photos are used to classify them according to age and gender, generating demographic data that will also be instantiated in the ontology. The goal is to demographically classify users to create a profile of people and help guide marketing strategies. At the end of this activity, all the information generated by the SMA sub-process is instantiated in the ontology. We use the predicted age to group users according to (70).

Figure 14 shows an example of an ontology graph focusing on the individual publication “tweet_1875669376425” belonging to the “Coalho” category. The publication had the sentiment classified as positive and was grouped into the sub cluster Coalho_Subcluster0 because it has the keyword “keyword_churrasco”. The user “user_4536323” who posted the “tweet_1875669376425” was classified as male and belongs to Generation X.

Figure 14 – Ontology graph for “tweet_1875669376425”.

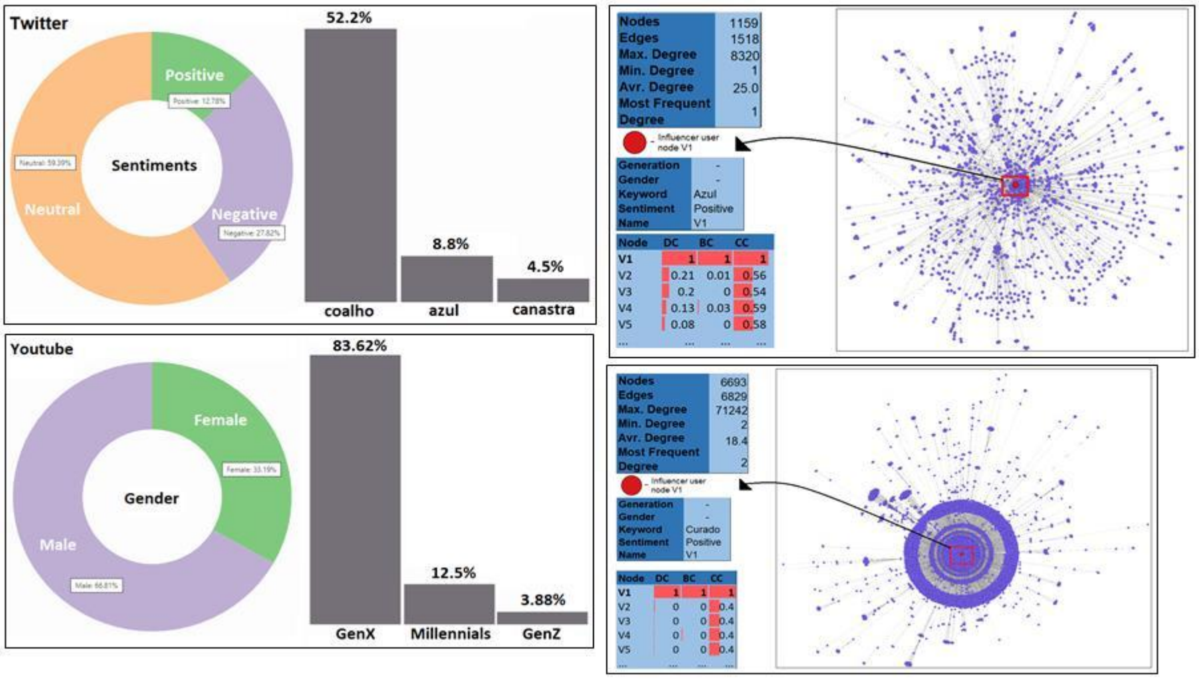


4.2.3 **Viewer Phase**

During the Viewer phase, the user can access the recommendations using the graphical interface, where she/he can view the most influential users in each OSN, or the most influential users, considering multiple OSN. The user can also access which products are the most mentioned in the OSN. The user can also get a list of users that has the potential to disseminate the product and can get the list of the “most” talked about related products. In section 5 an example of these recommendations will be presented. Figure 15 presents screenshots of a web application connected to IntelDigitalMarketing that presents the results of an ontology query for the for sentiments, keywords, genders, generations and influencers for mulitples ONSs, i.e., Twitter, Youtube and Instagram.

In the next section, we detail the IntelDigitalMarketing architecture, discussed the technologies used to implement the BPM-IntelDigitalMarketing phases.

Figure 15 – IntelDigitalMarketing´s web interface.



Nodes	1159
Edges	1518
Max. Degree	8320
Min. Degree	1
Avr. Degree	25.0
Most Frequent Degree	1

Generation	-
Gender	-
Keyword	Azul
Sentiment	Positive
Name	V1

Node	DC	BC	CC
V1	1	1	1
V2	0.21	0.01	0.56
V3	0.2	0	0.54
V4	0.13	0.03	0.59
V5	0.08	0	0.58
...	...	...	...

Nodes	6693
Edges	6829
Max. Degree	71242
Min. Degree	2
Avr. Degree	18.4
Most Frequent Degree	2

Generation	-
Gender	-
Keyword	Curado
Sentiment	Positive
Name	V1

Node	DC	BC	CC
V1	1	1	1
V2	0	0	0.4
V3	0	0	0.4
V4	0	0	0.4
V5	0	0	0.4
...	...	...	...

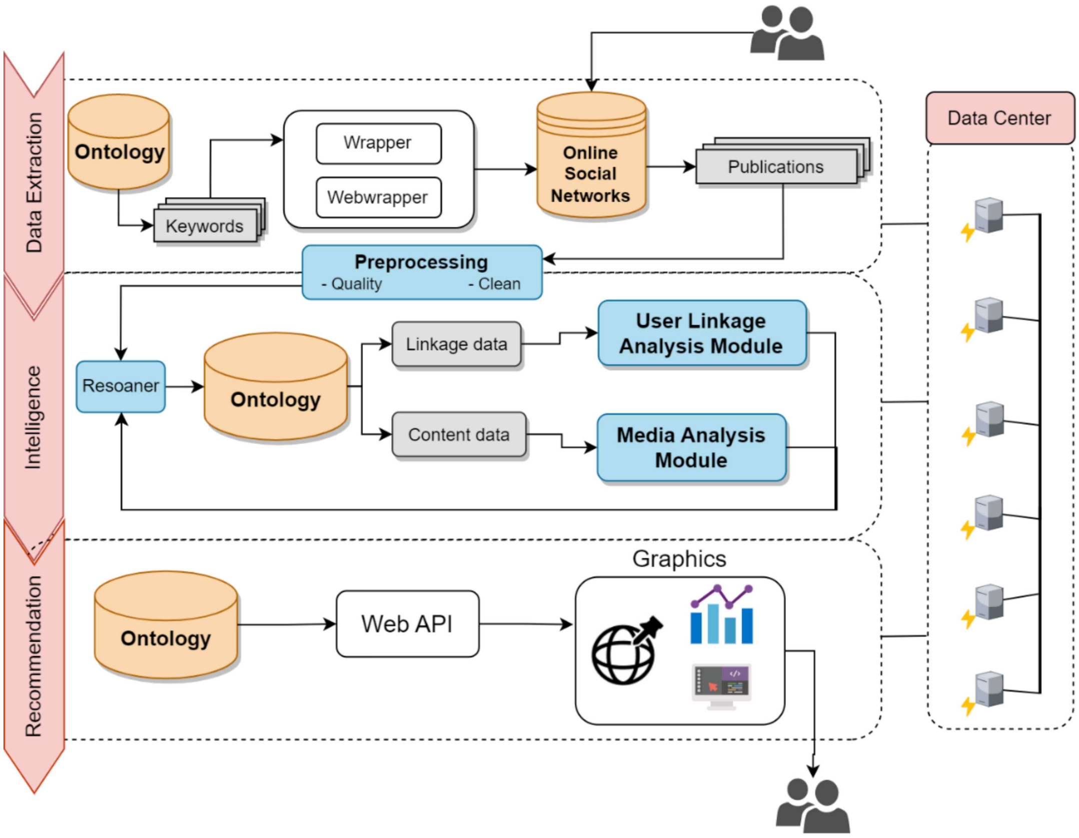
4.3 INTELDIGITALMARKETING ARCHITECTURE

As stated in the previous section, the IntelDigitalMarketing architecture is used to conduct the phases defined in BPM-IntelDigitalMarketing. Figure 16 presents a conceptual view of the IntellDigitalMarketing architecture. In the following sections, we discuss the details of the architecture layers.

To host the architecture, we use the computational infrastructure of the RePesq[17] laboratory located at UFJF, allowing optimized performance for data processing.

[17] https://www.repesq.ufjf.br/ (in Portuguese)

Figure 16 – IntelDigitalMarketing layered architecture



4.3.1 **Data Extraction Layer**

In this layer, the information from the OSNs is obtained through a translator (Data Extraction component in Figure 16), developed in Python[18], which contains an API for data extraction in the official OSNs´ repositories (Wrapper subcomponent in Figure 16). Currently, the architecture uses the official Twitter API[19], which pre-processes and stores the relevant information in the repository based on the domain terms specified in the ontology. The same occurs for the other OSNs mined. The

[18] https://www.python.org/

[19] https://developer.twitter.com/en/docs/twitter-api

architecture uses a webwrapper, developed using the requests[20] and selenium[21] libraries, to extract data from the web, such as from Instagram[22].

Also, the acquired data is separated into content and link data in the collection stage. In most cases, mined content is textual media[23]. Text analytics primarily deals with extracting meaningful data from unstructured data. The focus is to extract the necessary information, such as comments, search results, posts, tweets, blogs, logs, etc. Currently, the architecture uses the Natural Language Processing (NLP) algorithms from the NLTK[24] library for this pre-processing. The NLP techniques used were: (i) removing words from the publication text that make up a list of stop words from the NLTK library, which are meaningless words that are considered noise in the analyses; (ii) cleaning up and removing punctuation and numbers; (iii) cleaning up of repeated characters; (iv) removal of URLs; (v) obtaining tokenization of the publication text; and (vi) application of stemming.Extracted and processed data are instantiated in the ontology.

The ontology access is performed from a component that uses the OwlReady2[25] framework in Python. The ontology instantiation is executed using the Persistent World (PW) library, which allows the processing of large volumes of data in ontologies. This approach allows OwlReady2's optimized quadstore to be stored in an SQLite3[26] database. With the PW approach to load and execute inferences, the ontology has an optimized processing performance. Thus, using this approach, the ontology is divided into a model (TBox – only model description) and a complete ontology (Tbox + Abox, description of model and instances). From Tbox, a world is created to receive a limited number of instances and limit the reasoner's processing time (the architecture currently uses the Pellet[27] reasoner). After inference processing, this processed new world is loaded into the complete ontology for queries. PW allows the persistence of all the worlds of ontologies created and loaded.

[20] https://requests.readthedocs.io/en/latest/

[21] https://selenium-python.readthedocs.io/

[22] The webwrapper was used because the Instagram API has specific restrictions that make the data extraction process difficult.

[23] Most people express their views by speaking (audio and video) or writing (text)

[24] https://www.nltk.org/

[25] https://owlready2.readthedocs.io/en/v0.37/

[26] https://www.sqlite.org/

[27] https://github.com/stardog-union/pellet

4.3.2 **Intelligence Layer**

In this layer, the contents stored and processed in the ontology are provided, and also the processing of intelligent techniques to extract knowledge and assist recommendations. For this, this layer has two modules: (i) the module for analyzing connections between users, which models a network from the linked user data to analyze the relationships between them; and (ii) the collected media analysis module, where the textual and image content is classified.

4.3.2.1 User Linkage Analysis Module

In this layer, the connection data between users and content specified in the ontology are used to model a network of connected users. The module separates users according to their relationships, establishing communities. It also identifies which users are the most relevant in each OSN, and their interactions are characterized. In the architecture's current version, the OSN metric that measures how much users liked a post (such as Twitter Favorite or Instagram Likes) is used as interaction weights to model a weighted network. Results presented in (60) show that weighted networks play an important role in complex network techniques, and by assigning weights to relationships, we can improve predicting new relationships. Also, it represents the interaction's impact on OSN, considering that a post with a significant number of likes is an impactful post (61).

In this scenario, each community that belongs to a specific OSN creates a sub-network that is analyzed for its centrality and density. Network centrality measures how much it is centered on one or more key users or linkages, based on the number of relationships a user or linkage has within the network. This measure hints at how influence is distributed across a network. It is assumed that the number of relationships a user has, has a positive relationship with the power of influence that the user has over others. Network density measures the number of links within a network compared to the number of possible links, assuming all nodes are interconnected. This density is used to assess network coordination. So, the greater the network's density, the greater the potential for collaboration among identified users (62).

This module was developed using the Python language and the NetworkX[28] library to model and analyze the OSNs. We detail below the analyzes that are processed in the module.

- *Community Detection*. ONS users have their contacts added to their accounts from different aspects of their life, such as friends, work contacts, or persons with similar tastes. Thus, the processing is based on dividing contacts into groups to identify these different communities in the OSNs. The Louvain algorithm is one of the most used algorithms for community detection due to its speed and high modularity (63). Modularity values can vary from -1 to 1, and the higher the value, the better the community structure formed. We use the Community function of the NetworkX library, where the OSN is split into different potentially overlapping communities. The focus is to reduce inter-community edges and increase intra-community edges. Formally, the algorithm tries to maximize the network's modularity, or the fraction of edges that belong to a community minus the expected fraction of edges, if the edges were distributed randomly. Good communities in this context have a significant number of intra-community edges. Therefore, the greater the modularity, the denser communities with a high fraction of intra-community edges are detected. The Louvain method (63) consists of two phases: (i) the search for smaller communities, optimizing the modularity locally; and (ii) aggregating nodes from the same community and building a new network whose nodes are communities. These phases are repeated until a maximum of modularity is reached. Community detection is one of the most important research problems in SNA, where meaningful and cohesive subgroups are explored. Communities can be treated as potential functional units in complex network systems. In our work, the prediction score for future relationships is important, as it is a valuable source of similarity.
  - Figure 17 illustrates a network modeled in IntelDigitalMarketing using 1,000 tweets. With 1,928 nodes and 1,182 edges, the graph’s minimum degree is 1 and maximum is 2,299. The graph has disconnected nodes, which means they do not share any paths between them. Consequently, if a network contains sets of nodes that have no paths between them, the

[28] https://networkx.org/

network is also disconnected. One way to avoid this problem is to process metrics only for the nodes in the largest connected component. Another way is to separate the network using community detection and analyze them. Figure 18 shows the community graph, where the nodes are separated users according to their community color. We detected 770 communities with modularity of 0.72, where the most populous community C372 has 49 users. In general, communities have more than one user, but there are communities found by the analysis that are composed by only one user.

Figure 17 – Example of the generated graph of tweets using the networkx library

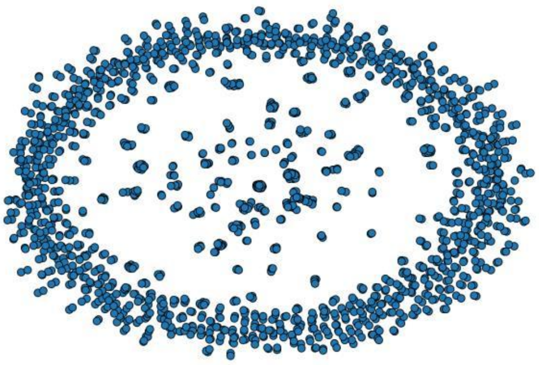
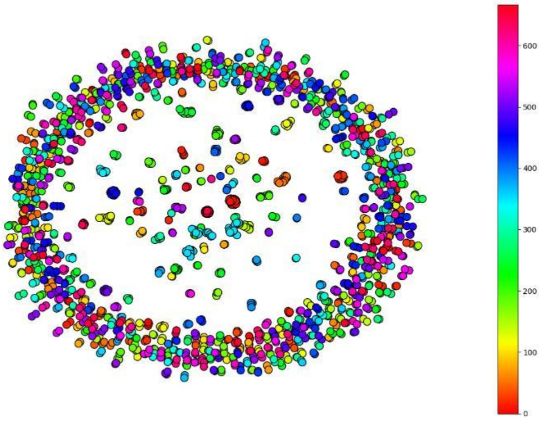


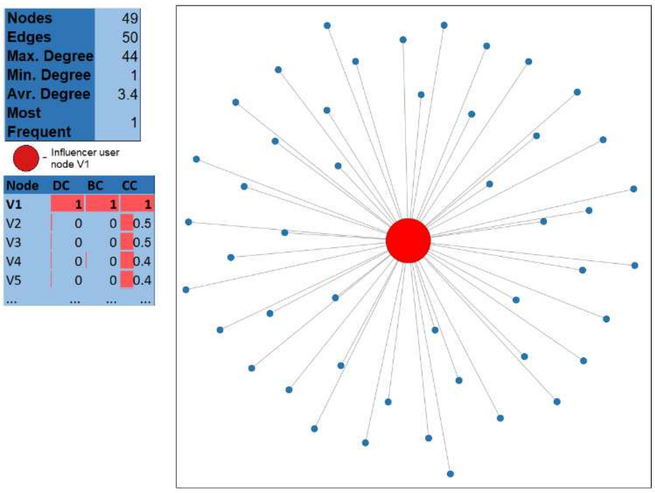
Figure 18 – Example of community graph generated from tweets using the networkx library.



- *Influence Analysis*. An Influencer is the most connected node (user) on the network. Each node in the network has a score representing its influence. This score is acquired through centrality algorithms, such as Page Rank and Betweenness Centrality (64), where nodes with a high score are the most influential. They have two main inputs: the graph and the weights on the graph's links. The Closeness Centrality algorithm (64) is also used, where nodes with a high score have the shortest distances to all other nodes. To visualize and quantify the relevance of the user in the network, the metrics Closeness (CC), Betweenness (BC), and Degree of Centrality (DC) (64) can be used. We used the DC measure, if an important user has many connections. However, the calculation of this measure does not consider the general structure of the network. For example, even if one node is connected to many others, it may not be a quick method of disseminating knowledge due to its position in the network. To consider the structure of the network, we used the CC measure. We used this measure to find users who are close to other people. A limitation of this measure is that it is generally restricted to networks with related components since nodes belonging to different components do not have a path between them. However, this does not apply to the analysis used in our work, as we separate the network into connected communities. Furthermore, we must consider that nodes that are not directly connected do not mean they cannot interact in the network. They can communicate through other nodes. Thus, we use the BC measure to find a user who is a bridge to other users.

- It is important to note that computing the centrality between all nodes in a graph involves computing the shortest paths between all pairs of nodes in this graph, which takes $O(|V||E| + |V|^{2}\log|V|)$ time using Brandes algorithm (65). Therefore, depending on the size of the network, it may not be feasible to use the BC metric. However, grouping the network into communities helps with this drawback by reducing the number of nodes in the BC calculation. To achieve these network metrics, we developed a module that uses the functions of the NetworkX library: (i) closeness_centrality, for the definition of the CC score; (ii) betweenness_centrality, to define the BC score; and (iii) degree, to define the GC score.
- Figure 19 shows an example of the subgraph corresponding to the most populous community emphasizing the most influential node. The subgraph's general information can also be seen in Figure 19, in which the subgraph is a connected network. The result of the influence analysis is a list with the centrality score for each node. In Figure 19, the red node has the highest score in all metrics. It represents the user with the greatest potential to influence the community. As it is a small network, it is easy to see the most influential user just from the structure of the network. However, in general, the network structure of the generated communities does not have this format and has more users, making it difficult to visualize the most influential user of the network.In this case, additional metrics must be used.

Figure 19 – Example of the most populous community generated graph using the networkx library.



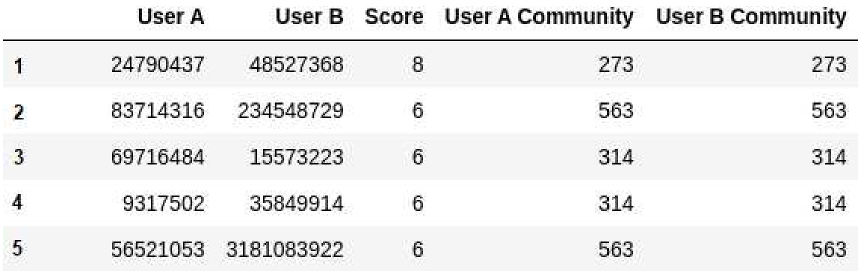
Nodes	49
Edges	50
Max. Degree	44
Min. Degree	1
Avr. Degree	3.4
Most Frequent	1

Node	DC	BC	CC
V1	1	1	1
V2	0	0	0.5
V3	0	0	0.5
V4	0	0	0.4
V5	0	0	0.4
...	...	...	...

- *Link Prediction*. Social media data is dynamic and evolving. Thus, a relevant issue in OSNs is the prediction of relationships, the analysis of similar social profiles, and its effect on the anticipation of relationships in the social network. The purpose of link prediction in IntelDigitalMarketing is to predict new relationships between users using their communities as a source of similarity in order to expand the network's relationships, contributing to intra-community and extra-community dissemination. We used a method based on the Soundarajan and Hopcroft method (59). For each pair of nodes u and v, the method calculates the number of common neighbors and a bonus for each common neighbor belonging to the same community as u and v.
    - As an example, Table 7 illustrates the five most scored (Score column) possible connections between User A and User B predicted for the entire network. Observing several predicted relationships between users of the same community is possible. We used the cn_soundarajan_hopcroft[29] function from the NetworkX library.

[29] https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_prediction.cn_soundarajan_hopcroft.html

Table 7 – Example of the table with results from the prediction links using the Pandas Image[30] library.



	User A	User B	Score	User A Community	User B Community
1	24790437	48527368	8	273	273
2	83714316	234548729	6	563	563
3	69716484	15573223	6	314	314
4	9317502	35849914	6	314	314
5	56521053	3181083922	6	563	563

4.3.2.2 Media Analysis Module

Social media analytics aims to know people's behavior so that we can create personalized content for everyone. Therefore, the techniques are developed to get a more detailed view of how data flows or how that data is related to user preference. Target media can be of any type. Therefore, we need to have different techniques for different media types. In general terms, four media types were analyzed: textual, audio, image, and video.

Currently, IntelDigitalMarketing analyzes textual and image media. This information is extracted from the ontology using SPARQL[31] queries. Then, different techniques are used in IntelDigital Marketing for processing these analyses.

- *Sentiment Analysis*. Sentiment analysis is the computational identification and classification of sentiments expressed as a written text about any specific topic, social network, or product review, among others (14). The aim is to identify sentiments in text media and classify them based on their polarity. Simple methods for sentiment analysis include word counting (the more a product is mentioned, the more it is supposed to be liked) or lists of positive and negative terms that can be counted when used, e.g. text media that mention a product. We used the list method to classify the media collected from the ONSs. The main functionality of this method is to classify data into entities and to recognize the class/entity to which a given entry belongs. If the prediction value tends to be categorized as yes/no, positive/negative, etc., it falls under the classification

[30] https://pypi.org/project/dataframe-image/

[31] https://www.w3.org/TR/rdf-sparql-query/

type problem in machine learning. To perform this classification of input data on entities and then establish relationships between these entities based on data, we use a dataset of positive, neutral, and negative terms to train a classifier using the supervised machine learning technique. We use the Kaggle[32] online platform. We can find the training dataset for the model used in[33]. This dataset was the most relevant considering the specific needs of IntelDigitalMarketing use for portuguese entrances. However, the dataset can be changed according to the domain.

- *Age/Gender Prediction*. Image analysis deals with extracting meaningful information from a dataset composed of images. The main applications of image analysis are facial recognition and motion analysis. IntelDigitalMarketing uses facial recognition on user profile photos that have been collected to classify users according to their age and gender. Age and gender prediction uses the model presented in (66), which combines custom facial recognition models based on deep learning to predict information. User demographics such as age and gender play an important role in marketing. They allow companies to enhance their services and segment the user. For this, we use the DeepFace[34] algorithm that combines two classification models where the input of both models is the user's photo. The output of the gender prediction model is the classification corresponding to the male and female gender only. Then the age model sorts the output with ages from 0 to 100. User profile pictures are made available by OSNs via URLs. To access the urls we use the python wget[35] library.
- *Topic Modeling.* Topic modeling is an unsupervised technique that analyzes the text of collected publications by assigning themes and organizing the publications into groups based on the assigned topics. Topic modeling attempts to group documents based on similar topics. Themes are used to provide consistent labels for exploring the collection of publications. The topics found can be used to support marketing, such as discovering user interests, detecting trends, or discovering hidden structures. IntelDigitalMarketing uses the Latent

[32] https://www.kaggle.com/

[33] https://www.kaggle.com/datasets/faelk8/portuguese-sentiment-analysis

[34] https://github.com/serengil/deepface

[35] https://pypi.org/project/wget/

Dirichlet Allocation (LDA) unsupervised machine learning technique (67). LDA takes documents as input and finds topics as output. Thus, the publications collected from the OSN are processed to find emerging topics. For this, we used an algorithm developed using the Python Scikitlearn library[36], using the GridSearchCV function to find the best model.

4.3.3 **Recommendation Layer**

In this layer, the results of the different analyzes are aggregated, evaluated, and recommended based on the semantic rules specified in the ontology. Using a semantic model, we reduce the impact of the significant problems that hinder the processing of traditional recommender systems (e.g., cold start, data dispersion, diversity, among others) and social recommender systems (e.g., heterogeneity data, data volume, and data volatility) (26).

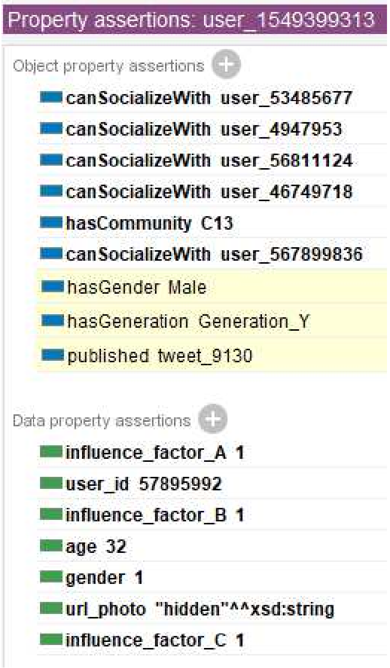
IntelDigitalMarketing uses an ontological model to represent the knowledge generated. While the folksonomy of keywords is statically defined to represent the content, the users semantic profiles are dynamic, They evolve based on their activities, behavior, and connections in the OSN. Therefore, the recommendation of relevant results is based on the results processed by the inference algorithm (Pellet reasoner) used in the ontology.

The outputs are made available from a dashboard (Figure 15), which allows the results analysis based on charts and through specific queries on the data made available. We can recommend information's about users and theirs content, i.e., the most influential user of the most populous community and the positives publications that he talk about. An example of a recommending result, based on the processing of semantic rules, is shown in Figure 20 (using Protégé[37] tool) for an individual user.

[36] https://scikit-learn.org/

[37] https://protege.stanford.edu/

Figure 20 – Example of an male, generation Y, influencer user recommendation (using the Protégé tool interface to present the instances details and the processing of inferences (in yellow))



Property assertions: user_1549399313

Object property assertions

canSocializeWith user_53485677
canSocializeWith user_4947953
canSocializeWith user_56811124
canSocializeWith user_46749718
hasCommunity C13
canSocializeWith user_567899836
hasGender Male
hasGeneration Generation_Y
published tweet_9130

Data property assertions

influence_factor_A 1
user_id 57895992
influence_factor_B 1
age 32
gender 1
url_photo "hidden"^^xsd:string
influence_factor_C 1

The data instantiated in the ontology is queried through an algorithm developed in Python that uses the SPARQL query language. We used the flask[38] and owlready2 libraries. The web interface was developed using the D3.js[39] library.

4.4 FINAL REMARKS

This chapter presented the IntelDigital Marketing solution, presenting the activities that must be done to use the solution, though the discussion of BPM-IntelDigitalMarketing and also detailed the IntelDigitalMarketing architecture and its implementation details. In the next chapter an evaluation of the solution is conducted.

[38]https://flask.palletsprojects.com/en/2.2.x/

[39] https://d3js.org/

5 EVALUATION

The Design Science methodology emphasizes the importance of an adequate evaluation. A Historical Research (18) is the recommended evaluation technique to answer explanatory questions when there is virtually no control over the events. The historical research advantage is that it does not depend on direct observations of the events. Instead, the study data sources rely on documents and artifacts as the primary sources of evidence.

Therefore, using Historical Research, we create knowledge about the use of technology. The researcher can evaluate if this technology meets the initially defined objectives, justifying the maintenance of the research. The knowledge obtained provides a basis for the refinements and the generation of new hypotheses to be investigated in future research.

5.1 PLANNING

As discussed in the Introduction of this dissertation, despite having great economic importance, the dairy derivatives sector needs more targeted dissemination strategies for a more specialized audience. In addition, knowing the consumer profile and market trends is strategic for the sector’s growth. To investigate if we can deal with this problem using our proposal, this evaluation considers the use of IntelDigitalMarketing to prospect the consumer profile of dairy products, including the prospection of market trends, helping producers and researchers to decide the best strategies for disseminating products. In this sense, we decided to mine the main available OSNs to verify the feasibility of the solution and answer the research questions.

We conducted this study using the infrastructure of NeNc group and REPESQ laboratories, at Federal University of Juiz de Fora with the help of researchers from the dairy derivatives and economics group from Embrapa Gado de Leite that identified the necessities for better marketing strategies. We also use the EMBRAPA- Gado de Leite computational infrastructure to host the web interface and help to process data.

The IntelDigitalMarketing was used to mine the main OSNs (Twitter, Instagram and YouTube, identified according to the literature mapping conducted) data. The scope of the study was defined through the GQM (Goal, Questions, and Metrics):

"**Analyze the use of the IntelDigitalMarketing solution from the point of view of producers and researchers of dairy derivatives in the context of data extracted from multiple OSNs**".

From this scope, the RQ was derived: **"How to foster Agribusiness Products by analyzing Multiple OSNs?"** As secondary Research Questions: **SRQ1**. What is the marketing effect of the most cited cheese among the communities of the different generations of users who talk about cheese on the OSN? **SRQ2**. Who are the biggest influencers of cheese at OSN?

Therefore, to help to answer these RQs, we formulated specific questions to be investigated in this study.

a) Which cheese is the most cited among communities?. According to the literature and our initial searches, cheese is the most cited dairy derivative. Therefore, specific searches considering this dairy derivative can help to detect dairy derivatives communities more precisely and identify specific products.
b) What is the feeling (positive or not) of users in the most populous cheese community concerning dairy derivatives? This question will help to know if dairy derivative positively impacts the community.
c) What is the profile of users who talk about cheese? This question helps to identify potential influencers to disseminate products or new products possibilities. To stratify this question, we also investigated: Which is the predominant generation? What is the predominant genre?
d) Who are the users who can disseminate dairy derivatives in other communities? The objective is to detect possible relationships between users using their communities as a reference. With this, it is possible to detect a possible relationship between users from different communities to spread the content across the network in an effective way.
e) Who are the biggest influencers interested in cheese at a specific OSN and what they talk about? The objective is to quantify the importance of each user within a community and, as a result, find users capable of disseminating content in the community. Also, find what else they talk about beyond cheese by searching the sub clusters of their publications.

To conduct the study, we used the IntelDigitalMarketing to access the Twitter data collection through its application programming interface (API). The tweets are collected through the Tweepy[40] library, which was implemented using a list of keywords related to dairy derivates as a filter. The result of this collection process was 428,195 tweets from Brazil's Portuguese language, tweeted in different parts of the world during the periods of 01/01/2021 – 01/29/2021.

Youtube was accessed through its native API YoutubeAPI. Instagram was access using a webwrapper. A total of 22,672 publications were collected from 03/12/2012 – 12/31/2020 and 178,869 comments from 03/12/2012 to 07/10/2021 on Instagram, and 1,706 videos published from 02/27/2021 – 11/ 06/2021 and 326,982 comments from 05/20/2021 – 06/19/2021 from Youtube.

To mine these OSNs, the RePesq computational infrastructure, together with EMBRAPA-Gado de Leite computational infrastructure were used. The data mined in these OSNs can be reached at [41].

Therefore, considering the steps specified in BPM-IntelDigitalMarketing (section 3.2), and the data mined on OSNs, the Artchee-O domain ontology was specified, based on a folksonomy[42] developed and validated by researchers of EMBRAPA-Gado de Leite. Before Artchee-O specification, a search was conducted to find existing ontologies that could be reused in this evaluation. Unfortunately, there was no ontology on this specific domain.

5.1.1 **ARTCHEE Ontology**

The ARTCHEE-O ontology aims to define and relate dairy product domain concepts and, from semantic rules, extract information and new relationships, helping to discover new connections between consumers of dairy products and related content on the web.

To develop the ontology, the following phases (73) was executed: (i) Specification, (ii) Conceptualisation, (iii) Formalisation, and (iv) Evaluation.

In the specification phase we identify the purpose, scope, implementation language, and intended End-Users. Therefore, ARTCHEE-O was developed to support IntelDigitalMarketing semantic processing (purpose), the scope was defined

[40] https://www.tweepy.org/

[41] https://github.com/nedsons/dissertacao/tree/main/data

[42] https://github.com/nedsons/dissertacao/blob/main/folksnomy.csv

considering the dairy derivatives domain and its customers (scope). The OWL 2.0 and SWRL languages were used as implementation languages. The intended users are producers and researchers interested in the dairy derivatives domain dissemination.

The conceptualization phase focused on organizing and structuring the semantic meaning of data. This phase was based on the folksonomy and on the relationships defined between terms. The formalization phase was done using the Protegé tool and OWL e SWRL languages. A set of rules was also defined to support the semantic processing of the terms and discover of new relationships between them.

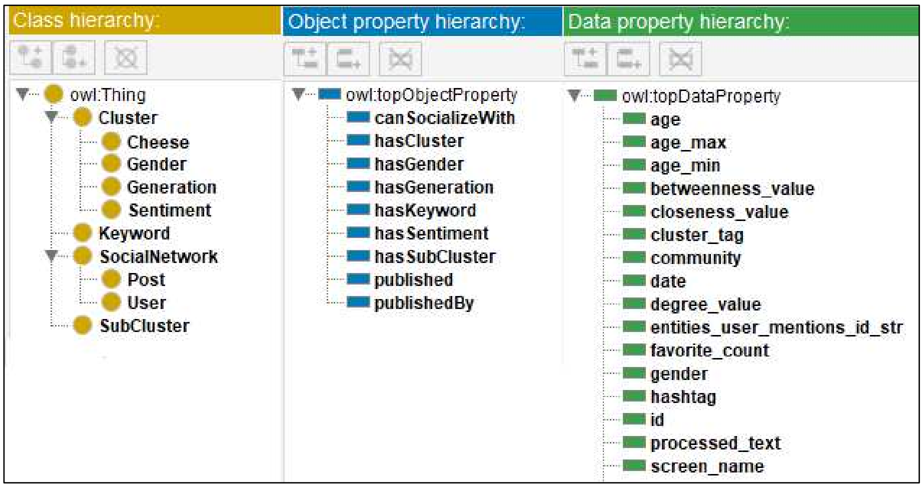
Finally, the evaluation phase consisting of carrying out a verification and validation. The verification step consists of verifying the correctness and validating the ontology. The correctness of the ontology is done through a verification process using the Pellet plugin reasoner on Protégé. The validation is a step to ensure that the ontology fulfills its purpose, though the answer of Competency Questions (CQ). The CQ specified for ARTCHEE-O are (Table 8):

Table 8 – Competence Questions.

CQ1	What are the publications that discuss about dairy derivative products in the OSNs?
CQ2	How do users feel about dairy derivative products?
CQ3	What is the generation and gender of users who talk about dairy derivative products?
CQ4	How are dairy derivative product categories divided and what are the subcategories?
CQ5	What are influencer communities?
CQ6	Who are the influencers in each community?

The queries (Table 9) provide specific answers to the Competence Questions (Table 8), validating the ARTCHEE-O ontology. With the ontology formalized and validated, ARTCHEE-O evolved by adding data without the dependent systems (IntelDigitalMarketing, for example) and processes being affected if changes occur. Figure 21 presents the main elements of ARTCHEE-O, using the Protégé tool interface.

Figure 21 – ARTCHEE-O's classes, object properties, and data properties. (Protégé tool screenshot)



Class hierarchy:

- owl:Thing
  - Cluster
    - Cheese
    - Gender
    - Generation
    - Sentiment
  - Keyword
  - SocialNetwork
    - Post
    - User
  - SubCluster

Object property hierarchy:

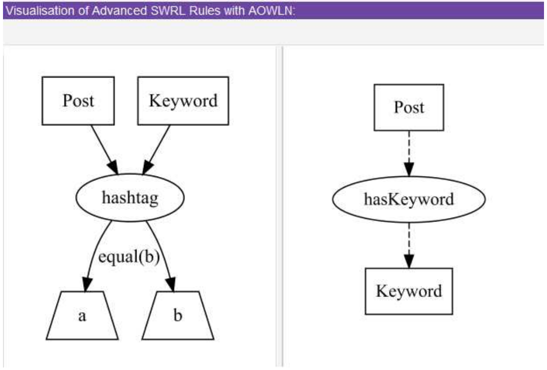
- owl:topObjectProperty
  - canSocializeWith
  - hasCluster
  - hasGender
  - hasGeneration
  - hasKeyword
  - hasSentiment
  - hasSubCluster
  - published
  - publishedBy

Data property hierarchy:

- owl:topDataProperty
  - age
  - age_max
  - age_min
  - betweenness_value
  - closeness_value
  - cluster_tag
  - community
  - date
  - degree_value
  - entities_user_mentions_id_str
  - favorite_count
  - gender
  - hashtag
  - id
  - processed_text
  - screen_name

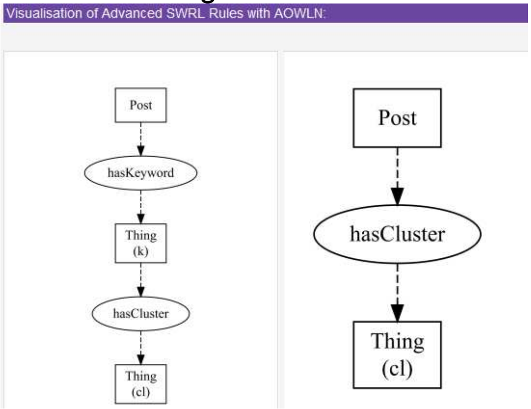
The ontology rules were developed using SWRL language (Figures 22-28), as stated before. First, the folksonomy concepts were instantiated into the ontology by relating each keyword to its category through the "hasCluster" object property. Them, the storage of publications in the ontology was done. **Rule1** (Figure 22) relates publications to keywords. According to its data property "hashtag", the object property "hasKeyword" can be inferred from the publication. **Rule2** (Figure 23) relates the publication to the category. In **Rule3** (Figure 24), the publication of the object property "publishedBy" can be inferred, indicating who owns it according to its data property "user_id". In **Rule4** (Figure 25), the object property "hasSentiment" can be inferred, which indicates the post’s sentiment. **Rule5** (Figure 26) relates publications to its subcategory, inferring the object property "hasSubCluster" according to its category. **Rule6** (Figure 27) relates users to their gender. Finally, **Rule7** (Figure 28) relates users to their generation according to predicted age.

Figure 22 – Rule 1



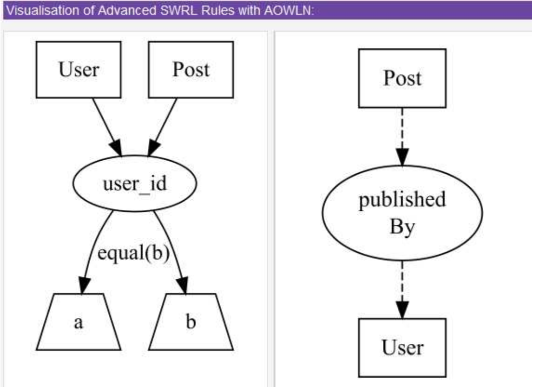
Rule1: Post(?po) ^ swrlb:equal(?a, ?b) ^ hashtag(?k, ?a) ^ Keyword(?k) ^ hashtag(?po, ?b) -> hasKeyword(?po, ?k)

Figure 23 – Rule 2



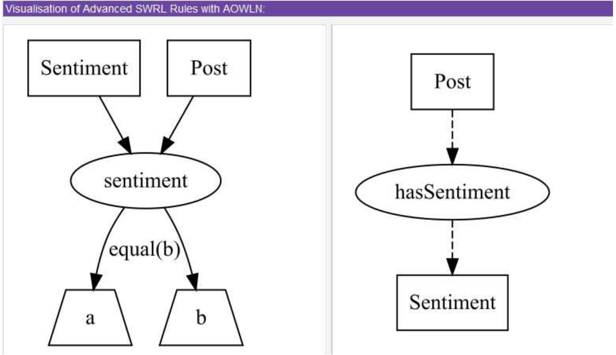
Rule2: Post(?po) ^ hasKeyword(?po, ?k) ^ hasCluster(?k, ?cl) -> hasCluster(?po, ?cl)

Figure 24 – Rule 3



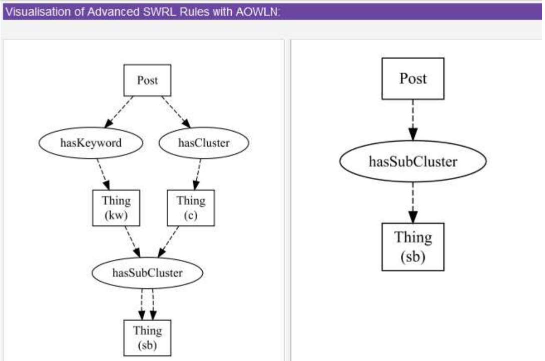
Rule3: Post(?po) ^ User(?us) ^ user_id(?us, ?a) ^ user_id(?po, ?b) ^ swrlb:equal(?a, ?b) -> publishedBy(?po, ?us)

Figure 25 – Rule 4



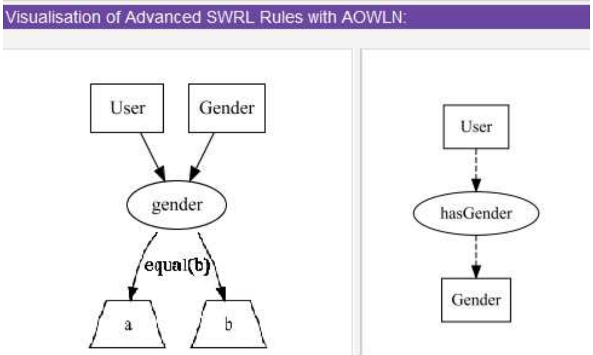
Rule4: Post(?po) ^ Sentiment(?se) ^ sentiment(?po, ?a) ^ sentiment(?se, ?b) ^ swrlb:equal(?a, ?b) -> hasSentiment(?po, ?se)

Figure 26 – Rule 5



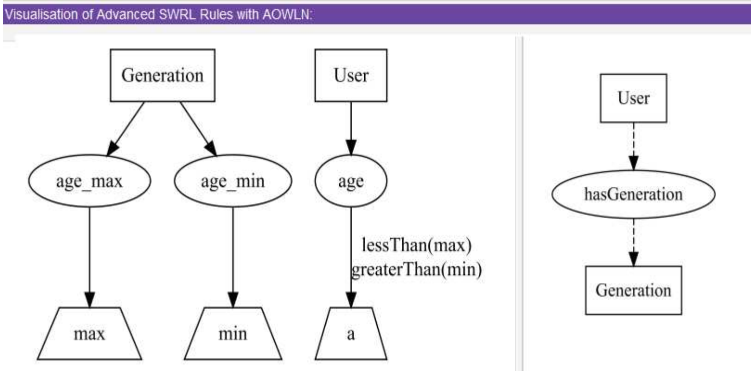
Rule5: Post(?po) ^ hasCluster(?po, ?c) ^ hasSubCluster(?c, ?sb) ^ hasKeyword(?po, ?kw) ^ hasSubCluster(?kw, ?sb) -> hasSubCluster(?po, ?sb)

Figure 27 – Rule 6



Rule6: Gender(?g) ^ User(?u) ^ gender(?g, ?a) ^ gender(?u, ?b) ^ swrlb:equal(?a, ?b) -> hasGender(?u, ?g)

Figure 28 – Rule 7



Rule7: Generation(?g) ^ User(?u) ^ age(?u, ?a) ^ age_max(?g, ?max) ^ age_min(?g, ?min) ^ swrlb:greaterThan(?a, ?min) ^ swrlb:lessThan(?a, ?max) -> hasGeneration(?u, ?g)

5.2 EXECUTION

During the evaluation execution, with ARTCHEE-O ontology and guided by secondary research questions SRQ1 and SRQ2 to answer RQ1, we collected a list of publications using as a filter the list of keywords related to cheese.

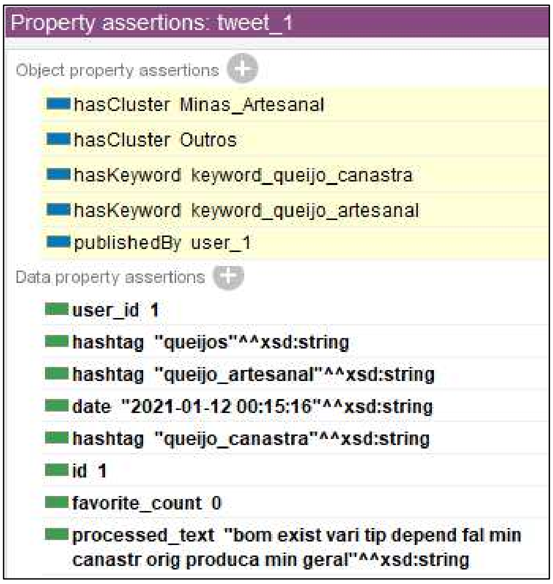
Using the ontology metadata, the result of the collection process was: (i) 428.195 tweets from Brazil's Portuguese language, tweeted in different parts of the world during the periods of 01/01/2021 – 01/29/2021;(ii) On Instagram 22,672 posts from 03/12/2012 – 12/31/2020 and 178,869 comments from 03/12/2012 – 07/10/2021 using only the keyword “artisanal cheese”. It was processed in this way to evaluate the ontology inference, relating the collected publications using a specific keyword with other keywords from the ontology; and (iii) 1,706 videos published between 02/27/2021 – 06/11/2021 and 326,982 comments posted between 05/20/2021 – 06/19/2021 on YouTube.

The result obtained from the processing of the reasoner on the ontological instances is shown in Figures 29, 30 and 31. According to the ontological metadata processing using Rule1 and Rule2, it was possible to classify some publications in more than one category in the three OSNs, i.e., the reasoner discovered new relationships “hasCluster” for the publication.

Figure 29 presents the property assertions focusing on the “tweet_1” ontological individual from Twitter. We can observe that the individual was classified into the category “Minas Artesanal” and “Outros”. In Figure 30, we have the individual “insta_1” from Instagram who has an inferred object property “hasCluster” for the category

"Outros" and "Coalho". Finally, Figure 31 shows the individual "video_1", where it is possible to observe the discovery (inference processing) of the "hasCluster" connection (objectProperty) for the categories "Artesanal Paulista" and "Outros". The "Other" category includes cheese-related keywords that do not fit into the other categories. Thus, at the end of data collection and processing, the ontology had a count of individuals equal to (i) 149,989 for Twitter; (ii) 329,221 for Instagram; e (iii) 203,033 for Youtube.

Figure 29 – Property assertions focusing on the individual "tweet_1" from Twitter. Protégé tool screenshot



Property assertions: tweet_1

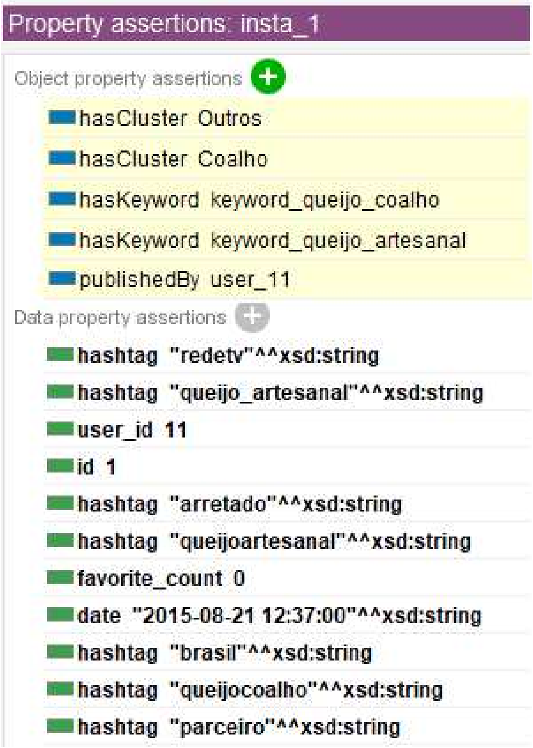
Object property assertions ⊕

- hasCluster Minas_Artesanal
- hasCluster Outros
- hasKeyword keyword_queijo_canastra
- hasKeyword keyword_queijo_artesanal
- publishedBy user_1

Data property assertions ⊕

- user_id 1
- hashtag "queijos"^^xsd:string
- hashtag "queijo_artesanal"^^xsd:string
- date "2021-01-12 00:15:16"^^xsd:string
- hashtag "queijo_canastra"^^xsd:string
- id 1
- favorite_count 0
- processed_text "bom exist vari tip depend fal min canastr orig produca min geral"^^xsd:string

Figure 30 – Property assertions focusing on the individual "insta_1" from Instagram. Protégé tool screenshot



Property assertions: insta_1

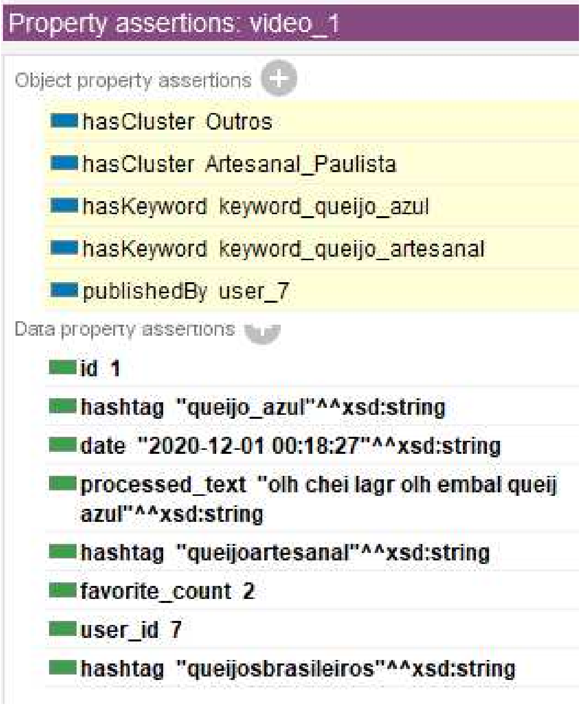
Object property assertions

hasCluster Outros
hasCluster Coalho
hasKeyword keyword_queijo_coalho
hasKeyword keyword_queijo_artesanal
publishedBy user_11

Data property assertions

hashtag "redetv"^^xsd:string
hashtag "queijo_artesanal"^^xsd:string
user_id 11
id 1
hashtag "arretado"^^xsd:string
hashtag "queijoartesanal"^^xsd:string
favorite_count 0
date "2015-08-21 12:37:00"^^xsd:string
hashtag "brasil"^^xsd:string
hashtag "queijocoalho"^^xsd:string
hashtag "parceiro"^^xsd:string

Figure 31 – Property assertions focusing on the individual "video_1" from Youtube. Protégé tool screenshot



Property assertions: video_1

Object property assertions

hasCluster Outros
hasCluster Artesanal_Paulista
hasKeyword keyword_queijo_azul
hasKeyword keyword_queijo_artesanal
publishedBy user_7

Data property assertions

id 1
hashtag "queijo_azul"^^xsd:string
date "2020-12-01 00:18:27"^^xsd:string
processed_text "olh chei lagr olh embal queij azul"^^xsd:string
hashtag "queijoartesanal"^^xsd:string
favorite_count 2
user_id 7
hashtag "queijosbrasileiros"^^xsd:string

Besides, the ontology also identified synonyms and spelling errors based on the specified rules. For example, if the keyword "queijo coalho" were misspelled for "queijo coalio", it is possible to mix it by assigning several data properties hashtag to the individual “keyword_queijo_coalho”. An example of a synonym can be seen in Figure 30, where the keyword “queijocoalho”, attributed to the individual “keyword_queijo_coalho”, was cited in the publication “insta1”. With the data collected from the OSNs instantiated in the ontology, IntelDigitalMarketing accessed the ontology to process the SMA, considering the content of the publications. From this analysis, sub clusters, sentiments, gender, and generation were discovered.

To help answer RQ, some specific queries were processed (Table 9). Using the SPARQL CS1 query (Table 9), IntelDigitalMarketing extracted from the ontology all publications separated by OSN to classify them according to specific sentiments (positive, negative, neutral). The classification was done through the trained model that adds the data property “sentiment” to the individual corresponding to the publication. An example of the object property “hasSentiment” inferred for the individual insta_1 can be seen in Figure 32.

After this sentiment classification, IntelDigitalMarketing could process the CS6 query (Table 9), generating a graph with the result of the analysis. The graph can be seen in Figure 33 for Twitter, Figure 35 for Instagram and Figure 34 for Youtube, where the pie chart with the result of the sentiment analysis and the bar chart with the top three keywords.

On Twitter, the dominant sentiment was “neutral”, with ~59% return, and the most cited keyword was “coalho”, which belongs to the “Coalho” category. On Instagram, the result was ~81% for “neutral” sentiment, with “positive” (~10%) and “negative” (~9%) sentiments being close to each other, with a difference of ~1%. The three most cited keywords of folksonomy on Instagram, filtering the keyword "queijo artesanal", which consequently is the most cited keyword, as it was the only one used in the OSN extraction, were the words “canastra” (~26 %), “queijos artesanais” (~24%) and “curado” (~11%), the first and second belonging to the “Minas Artesanal” category, and the third to the “outros” category.

On Youtube, IntelDigitalMarketing presented “neutral” as the dominant sentiment with ~65%, and the most cited keyword was “da marisa”, which belongs to the category “Minas Artesanal”.

By ranking publications classified according to sentiment and CS6 query, IntelDigitalMarketing recommended publications that helped identify trends, meeting Functional Requirement 3 (FR3). In addition, the ontology helped IntelDigitalMarketing in supporting the sharing of knowledge generated with other applications, meeting Functional Requirement 4 (FR4)[43].

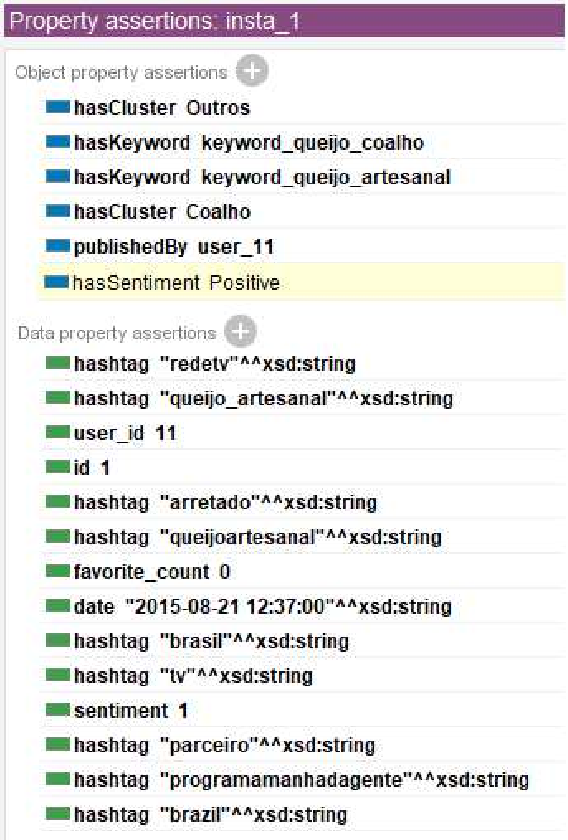
Table 9 – SPARQL queries

Indice	SPARQL Query	Parameter
CS1	SELECT ?Date ?Post ?processed_text WHERE { ?Post a ont:Post . ?Post ont:date ?Date . ?Post ont:processed_text ?processed_text FILTER (!EXISTS { ?Post ont:hasSentiment ?Sentiment})}	-
CS2	SELECT ?Date ?Post ?processed_text WHERE { ?Post a ont:Post . ?Post ont:date ?Date . ?Post ont:processed_text ?processed_text FILTER (EXISTS { ?Post ont:hasKeyword ont:<keyword_parameter> }) }	keyword_parameter
CS3	SELECT ?User WHERE { ?User a ont:User FILTER (!EXISTS { ?User ont:age ?age }) }	-
CS4	SELECT ?source ?target ?weight WHERE { ?Post a ont:Post . ?Post ont:entities_user_mentions_id_str ?target . ?Post ont:favorite_count ?weight . ?Post ont:user_id ?source }	-
CS5	SELECT ?Post ?Palavra ?Comunidade WHERE {?Post a ont:Post . ?Post ont:hashtag ?Palavra . ?Post ont:publishedBy ?user . ?user ont:community ?Comunidade FILTER (?Comunidade = <community_parameter>)}	community_parameter
CS6	SELECT * {{?Post a ont:Post . ?Post ont:date ?Date . ?Post ont:hasSentiment ?Sentiment . ?Post ont:hasKeyword ?Keyword} UNION {?Post a ont:Post . ?Post ont:publishedBy ?User . ?User ont:hasGeneration ?Generation . ?User ont:hasGender ?Gender}}	-
CS7	SELECT ?User1 ?User2 WHERE { ?User1 ont:canSocializeWith ?User2 }	-

[43] For example, APIs can query the ontology through SPARQL, which is a standardized language for Resource Description Framework (RDF) graph query and is a machine-understandable language.

CS8	SELECT ?id_origem ?id_destino ?peso ?community WHERE { ?Post a ont:Post . ?Post ont:entities_user_mentions_id_str ?id_destino . ?Post ont:favorite_count ?peso . ?Post ont:user_id ?id_origem . ?Post ont:publishedBy ?user_a . ?user_a ont:community ?community FILTER (?community = <community_parameter>) }	community_parameter
CS9	SELECT ?User WHERE { {?User ont:betweenness_value 1 . ?User ont:community ?Community} UNION {?User ont:closeness_value 1 . ?User ont:community ?Community} UNION {?User ont:degree_value 1 . ?User ont:community ?Community}}	-

Figure 32 – Example of “hasSentiment” object property, inferred to an individual. Protégé tool screenshot



Property assertions: insta_1

Object property assertions

hasCluster Outros

hasKeyword keyword_queijo_coalho

hasKeyword keyword_queijo_artesanal

hasCluster Coalho

publishedBy user_11

hasSentiment Positive

Data property assertions

hashtag "redetv"^^xsd:string

hashtag "queijo_artesanal"^^xsd:string

user_id 11

id 1

hashtag "arretado"^^xsd:string

hashtag "queijoartesanal"^^xsd:string

favorite_count 0

date "2015-08-21 12:37:00"^^xsd:string

hashtag "brasil"^^xsd:string

hashtag "tv"^^xsd:string

sentiment 1

hashtag "parceiro"^^xsd:string

hashtag "programamanhadagente"^^xsd:string

hashtag "brazil"^^xsd:string

Figure 33 – Twitter sentiment pie chart and top three keywords bar chart presented using the IntelDigital Marketing web interface

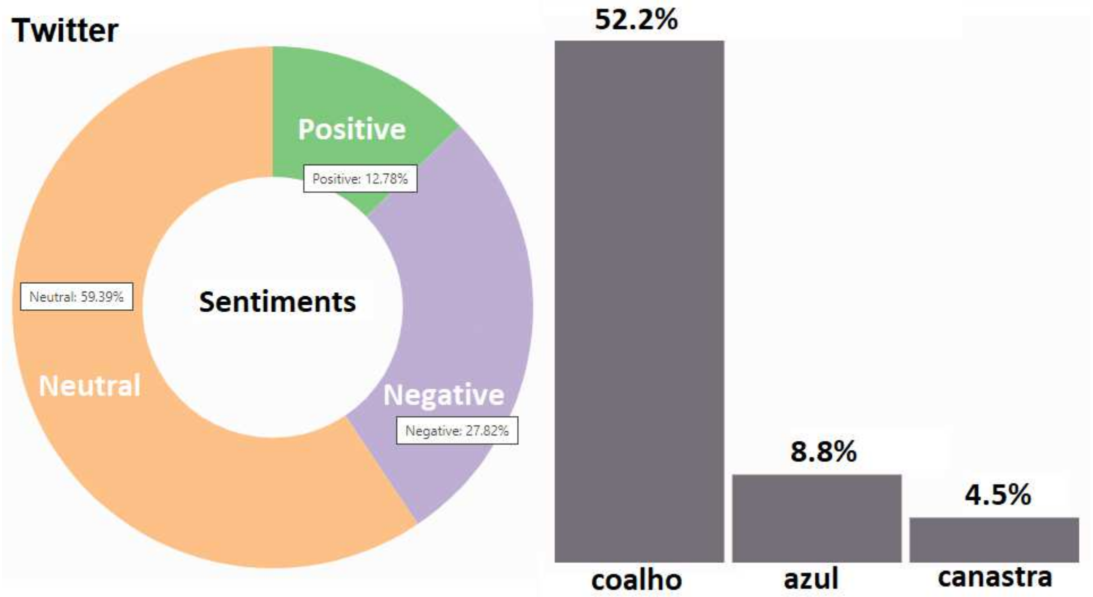


Figure 34 – Youtube sentiment pie chart and top three keywords bar chart presented using the IntelDigital Marketing web interface

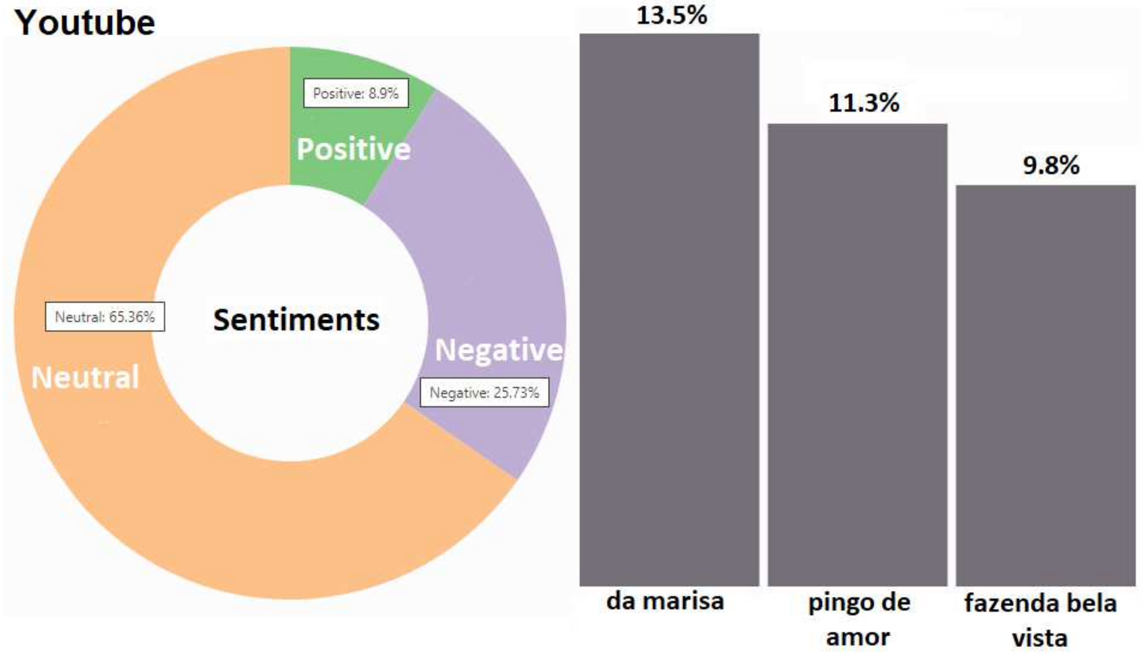


Figure 35 – Instagram sentiment pie chart and top three keywords bar chart presented using the IntelDigital Marketing web interface

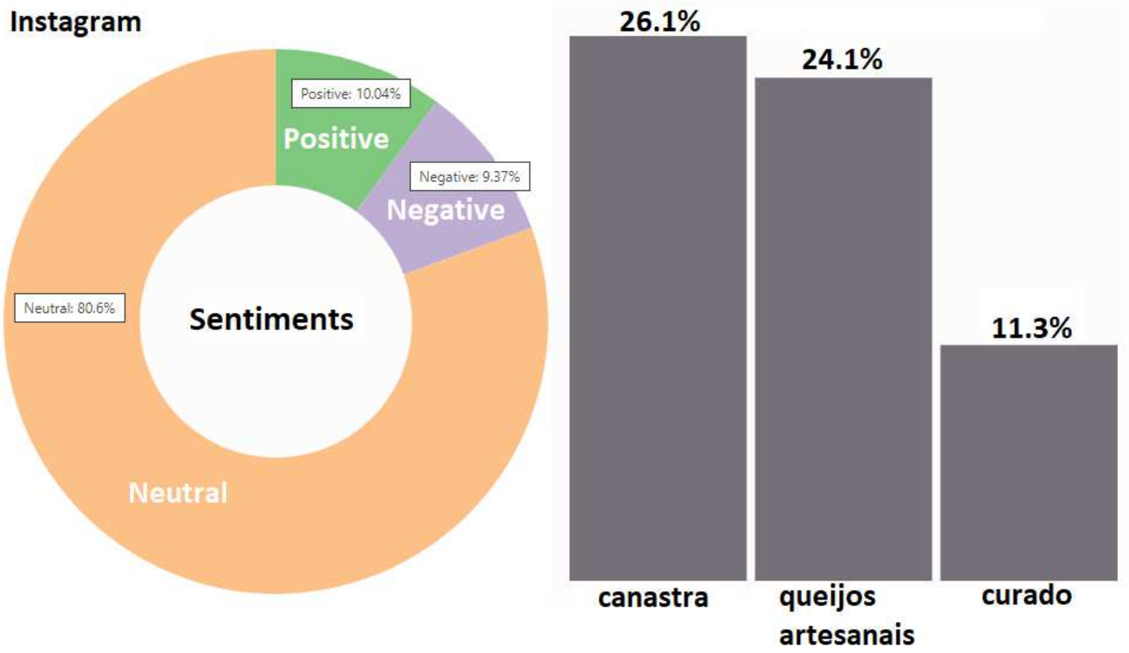
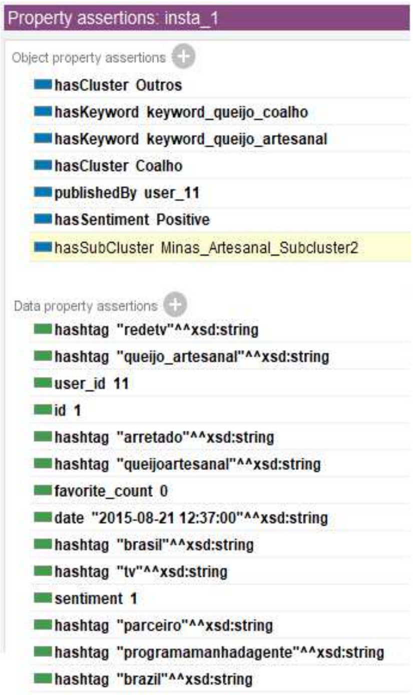


Figure 36 –“hasSubCluster” object property inference for “insta_1” individual. Protégé tool screenshot



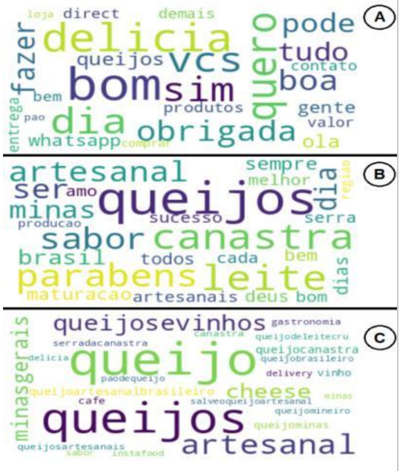
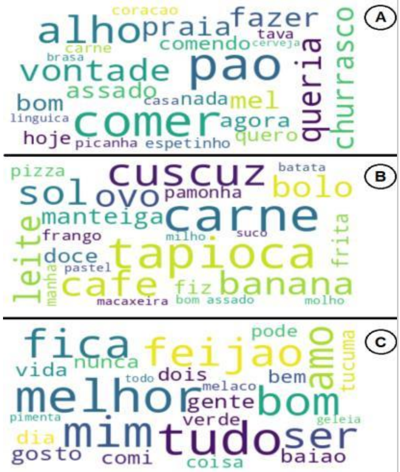
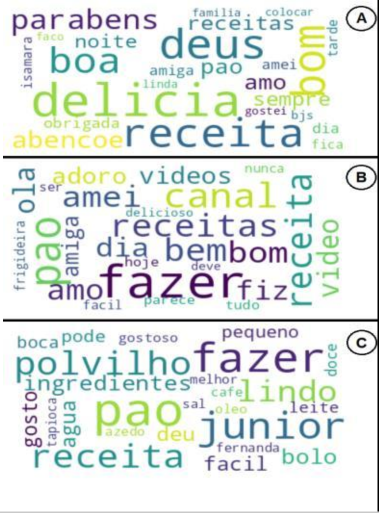
Property assertions: insta_1

Object property assertions

- hasCluster Outros
- hasKeyword keyword_queijo_coalho
- hasKeyword keyword_queijo_artesanal
- hasCluster Coalho
- publishedBy user_11
- hasSentiment Positive
- hasSubCluster Minas_Artesanal_Subcluster2

Data property assertions

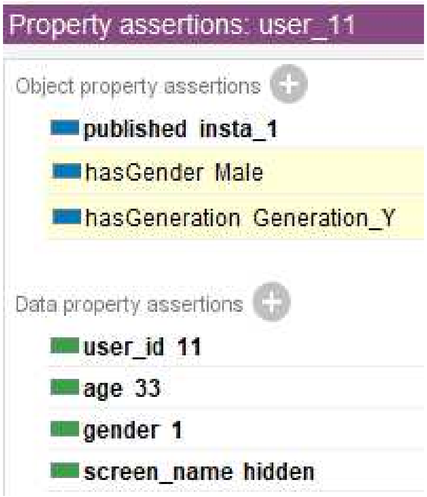
- hashtag "redetv"^^xsd:string
- hashtag "queijo_artesanal"^^xsd:string
- user_id 11
- id 1
- hashtag "arretado"^^xsd:string
- hashtag "queijoartesanal"^^xsd:string
- favorite_count 0
- date "2015-08-21 12:37:00"^^xsd:string
- hashtag "brasil"^^xsd:string
- hashtag "tv"^^xsd:string
- sentiment 1
- hashtag "parceiro"^^xsd:string
- hashtag "programamanhadagente"^^xsd:string
- hashtag "brazil"^^xsd:string

Figure 37 – (A) Sub Topic A; (B) Sub Topic B; and (C) Sub Topic C for the Instagram keyword “queijo artesanal”	Figure 38 – (A) Sub Topic A; (B) Sub Topic B; and (C) Sub Topic C for the Twitter word “queijo coalho”.	Figure 39 – (A) Sub Topic A; (B) Sub Topic B; and (C) Sub Topic C for the keyword “da marisa” from Youtube
		

With SPARQL CS2 (Table 9), we added the most cited keyword as a parameter to the query, IntelDigitalMarketing extracted texts from publications for subtopic modeling and the sub clusters found were instantiated in the ontology. From the inference processing using Rule5, it was possible to infer new “hasSubCluster” relationships for the individual. Figure 36 shows the object property inferred between individual "insta_1" and the sub cluster "Minas_Artesanal_Subcluster2". Figures 37, 38 and 39 show the wordclouds generated with the 25 most frequent keywords of each sub-topic, (returned from the ontology) where the main topic is the category of the most cited cheese found on Twitter (Figure 38), Instagram (Figure 37) and YouTube (Figure 39). In Figures 37, 38 and 39, each letter (A, B and C) corresponds to the wordcloud of a subgroup where the word size is proportional to its citation frequency.

Through CS3 (Table 9), IntelDigitalMarketing also extracted each user's profile pictures to demographically classify users. IntelDigitalMarketing use the information from CS3 to collect the user’s data from the user’s OSN database. The result of this analysis was stored in the ontology in the data property “age” and “gender”. Figure 40 shows the properties assertions of “user_11” with the “hasGender” and “hasGeneration” inferred. Using the CS6 query, IntelDigitalMarketing can generate the following information for each OSN: (i) Figure 41 for Twitter, which shows the pie chart with the percentage of each gender. With ~77%, the male gender was dominant. Figure 41 also shows the bar chart with the percentage of generations, where the dominant generation was “GenX” with ~85%; (ii) Figure 43 for Instagram, where the pie chart with the percentage of each gender was significantly more balanced than the other OSNs, with ~58% the male gender was dominant. Figure 43 also shows the bar chart with the percentage of generations, where the dominant generation was “GenX” with ~86%; and finally, (iii) Figure 42 for Youtube, where the gender pie chart shows men with ~67% and the bar chart with the percentage of generations, with the dominant generation “GenX” with ~84%.

Figure 40 – “hasGender” and “hasGeneration” object properties inferred for “user_11”. Protégé tool screenshot



Property assertions: user_11

Object property assertions

published insta_1

hasGender Male

hasGeneration Generation_Y

Data property assertions

user_id 11

age 33

gender 1

screen_name hidden

Figure 41 – Twitter gender pie chart and generation bar chart presented using the IntelDigital Marketing web interface.

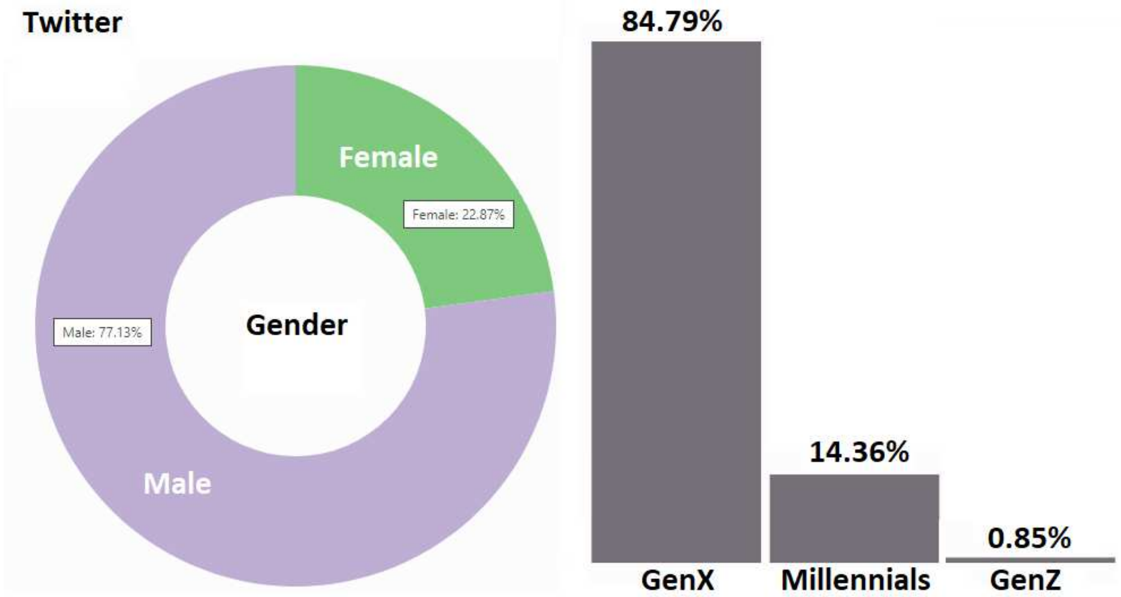


Figure 42 – Youtube gender pie chart and generation bar chart presented using the IntelDigital Marketing web interface.

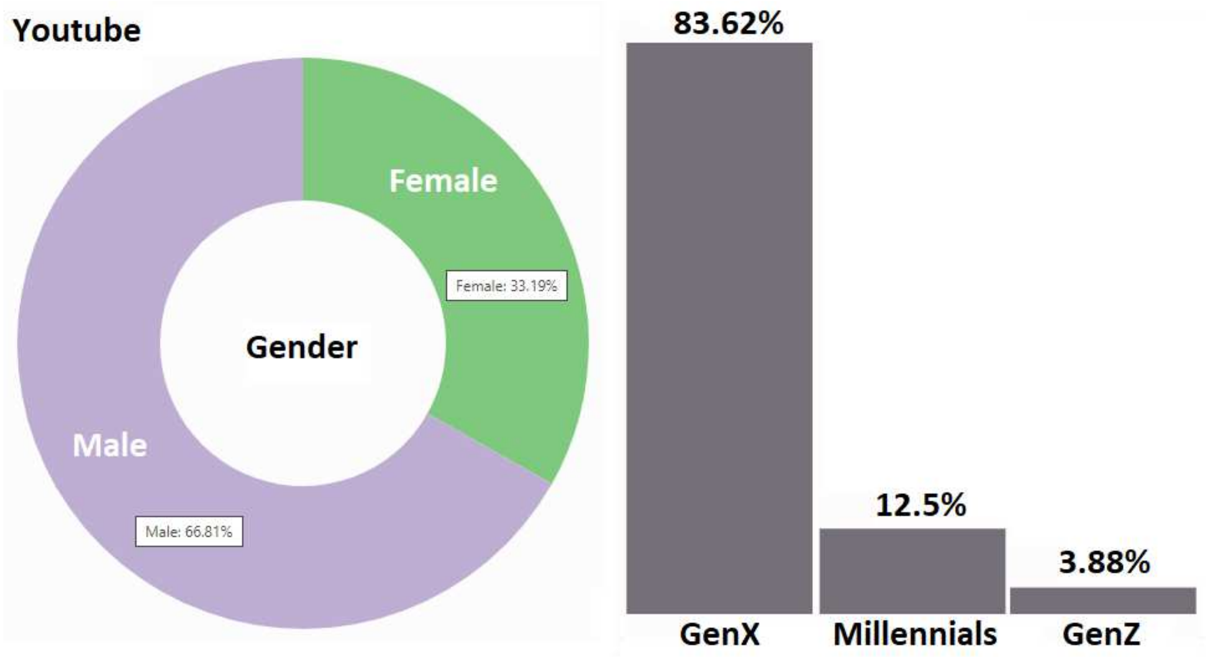
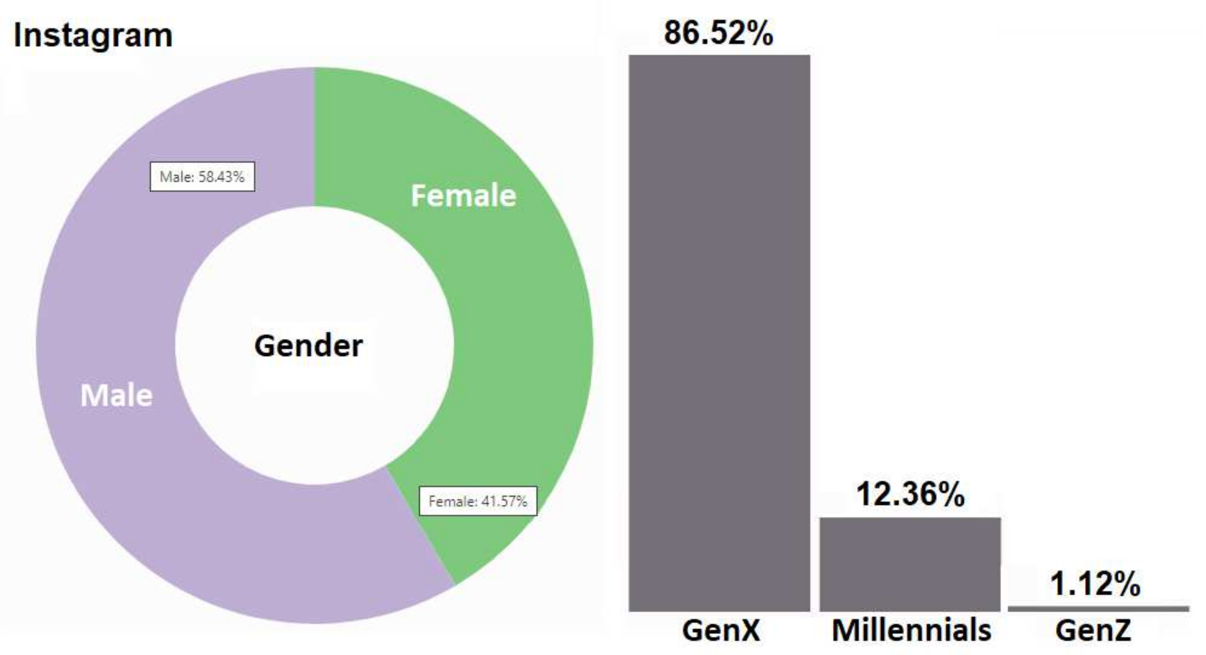


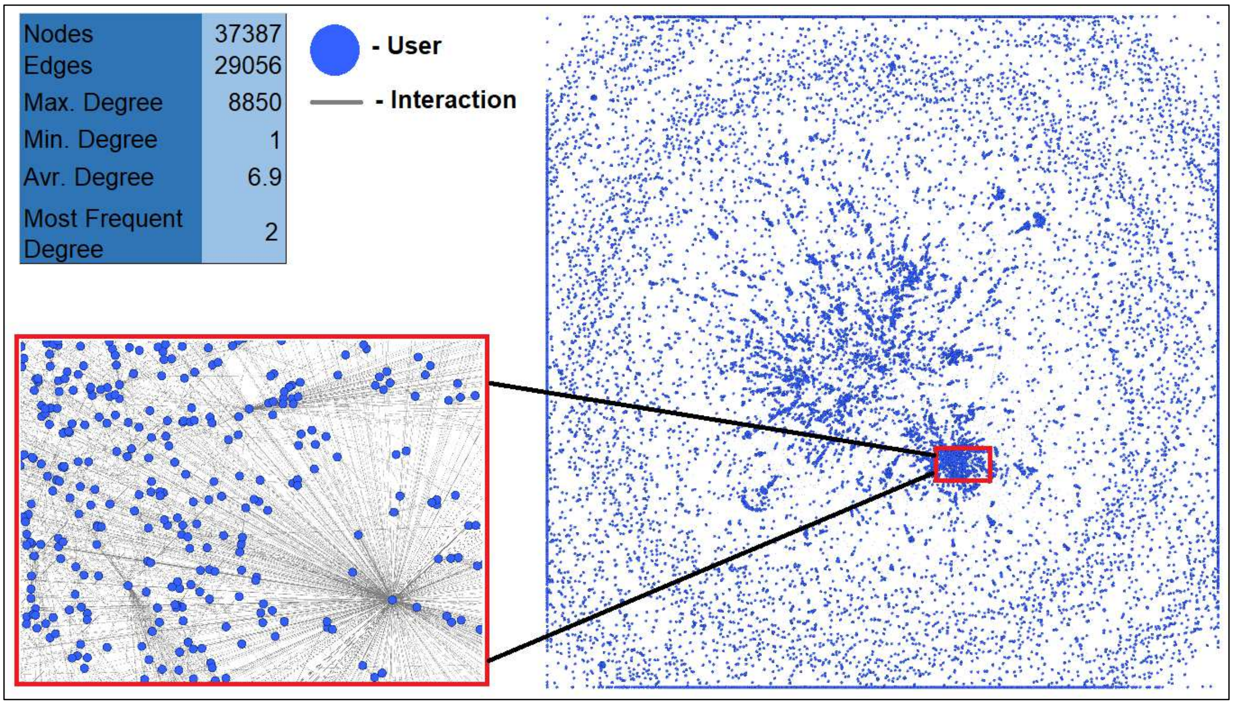
Figure 43 – Instagram gender pie chart and generation bar chart presented using the IntelDigital Marketing web interface.



The CS4 query (Table 9) extracted the interactions between users found in the collected publications, allowing the user network generation. The CS4 query extracts the Twitter mention interactions, which is when a user mentions the other in the post, and the Instagram and YouTube comment interactions (when a user comments on another user's post). The generated graphs can be seen along with their general information in Figure 44 for Twitter, Figure 45 for Instagram and Figure 46 for YouTube. The three OSNs present disconnected networks. Twitter and Youtube presented the minimum degree of the graph equal to 1, and Instagram equal to 2.

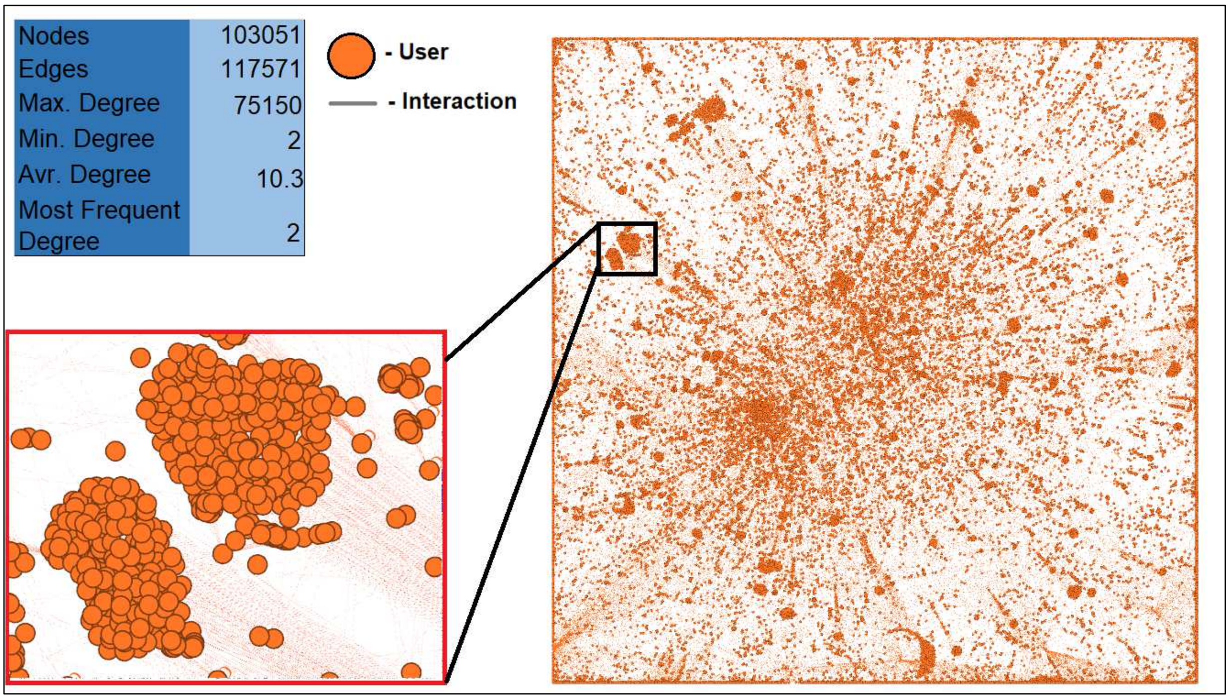
In this evaluation, it is important do emphasize, considering the processing cost to generate this graph, which would bring a delay in the presentation of the graph and its manipulation using the IntelDigitalMarketing web interface, Figures 42-44 were generated using the GEPHI tool, from a csv file with data returned from CS4 query. In future IntelDigitalMarketing versions, we will improve this functionality.

Figure 44 – Twitter graph. Gephi tool screenshot.



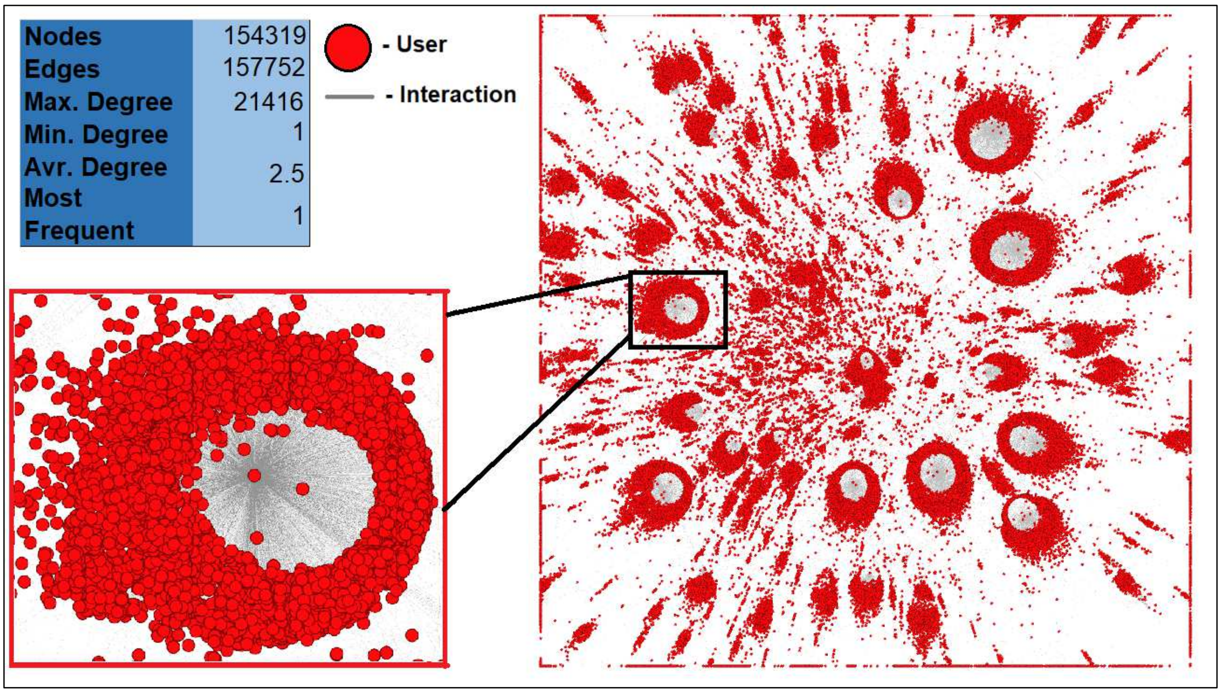
Nodes	37387
Edges	29056
Max. Degree	8850
Min. Degree	1
Avr. Degree	6.9
Most Frequent Degree	2

Figure 45 – Instagram’s graph. Gephi tool screenshot.



Nodes	103051
Edges	117571
Max. Degree	75150
Min. Degree	2
Avr. Degree	10.3
Most Frequent Degree	2

Figure 46 – Youtube’s graph. Gephi tool screenshot.

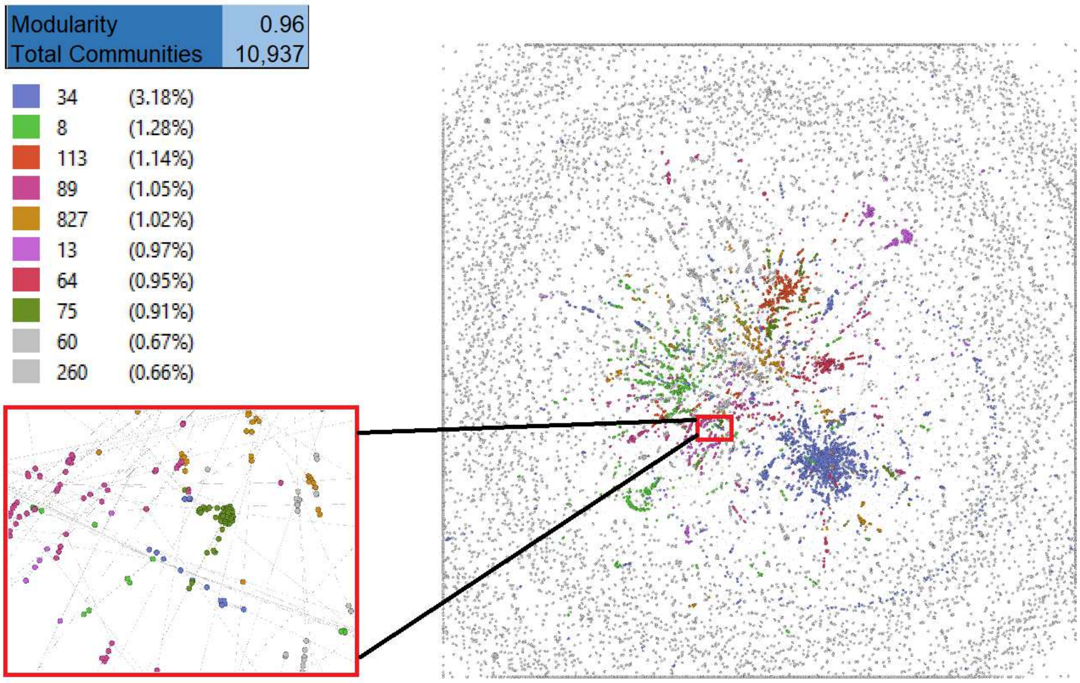


Nodes	154319
Edges	157752
Max. Degree	21416
Min. Degree	1
Avr. Degree	2.5
Most Frequent	1

Analyzing this number of nodes requires high computational costs. To analyze only connected networks and reduce the computational cost of the analysis and presentation of the results in the web interface, IntelDigitalMarketing divided the graph into user communities, as described in Section 4.2.1. Figures 47, 48 and 49 show the OSNs´ graphs where the nodes are users with colors corresponding to a community.

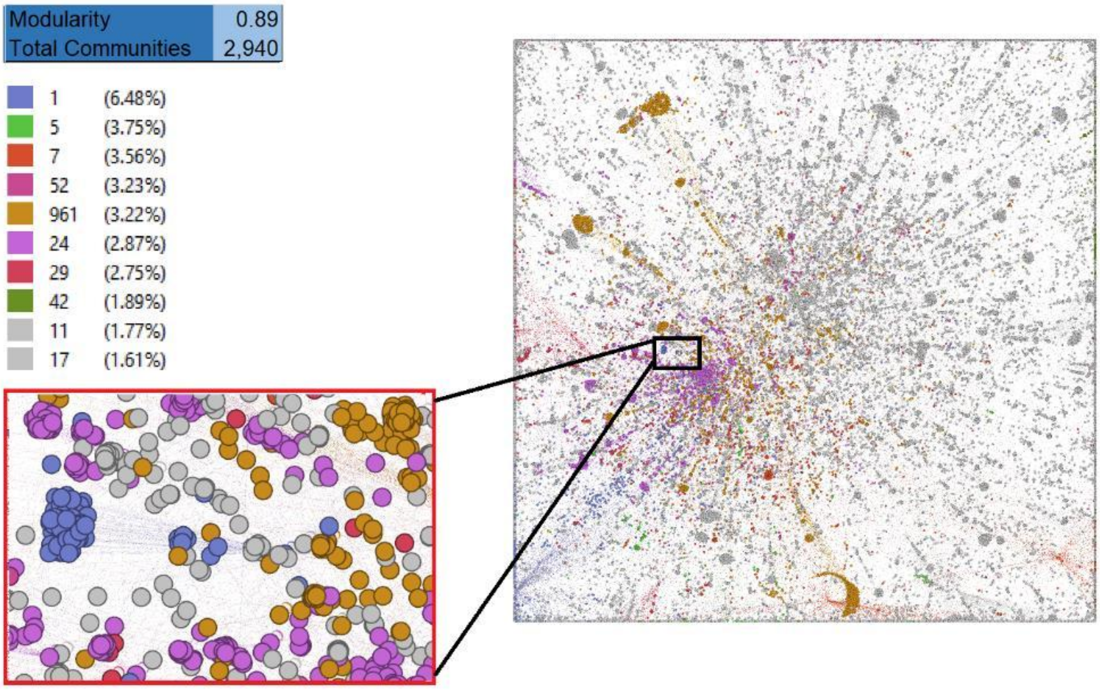
IntelDigitalMarketing detected 10,937 communities with a modularity of 0.96 on Twitter (Figure 47), 2,940 communities with a modularity of 0.89 on Instagram (Figure 48) and 817 communities with a modularity of 0.90 on Youtube (Figure 49). After detection, the community data were instantiated into the ontology in the "user" class, through the data property “community”. Figures 47-49 were generated using the Gephi tool, considering the processing costs.

Figure 47 – Twitter communities’ graph



Modularity	0.96
Total Communities	10,937

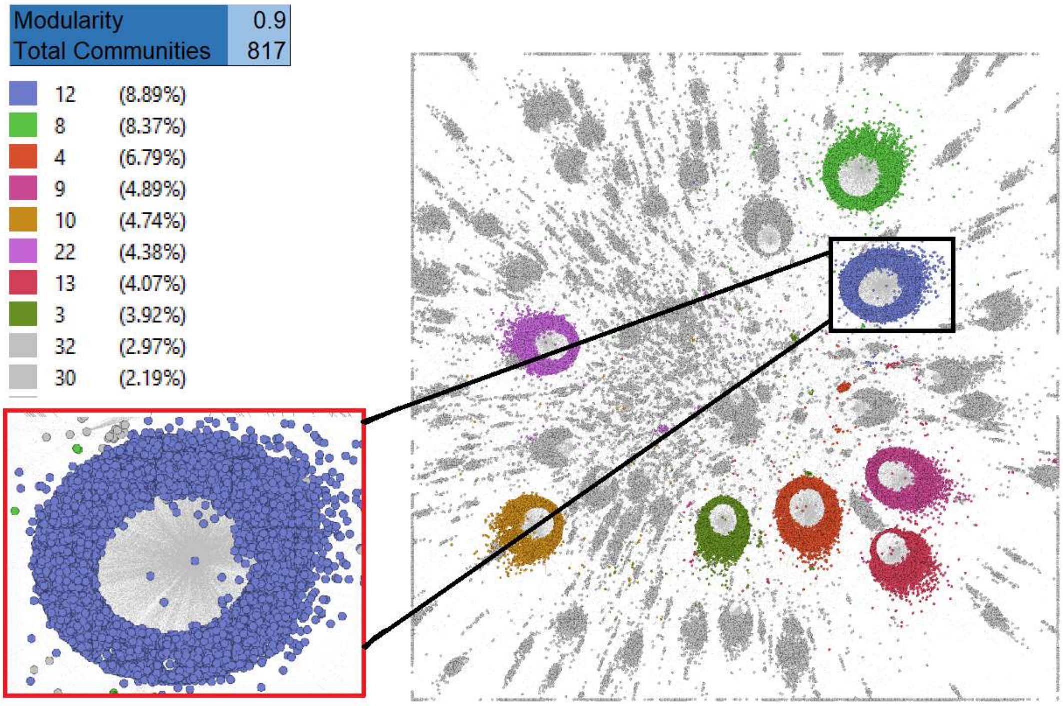
Figure 48 – Instagram communities graph



Modularity	0.89
Total Communities	2,940

🟦	1	(6.48%)
🟩	5	(3.75%)
🟥	7	(3.56%)
🟪	52	(3.23%)
🟧	961	(3.22%)
🟪	24	(2.87%)
🟥	29	(2.75%)
🟩	42	(1.89%)
⬜	11	(1.77%)
⬜	17	(1.61%)

Figure 49 – Youtube communities' graph

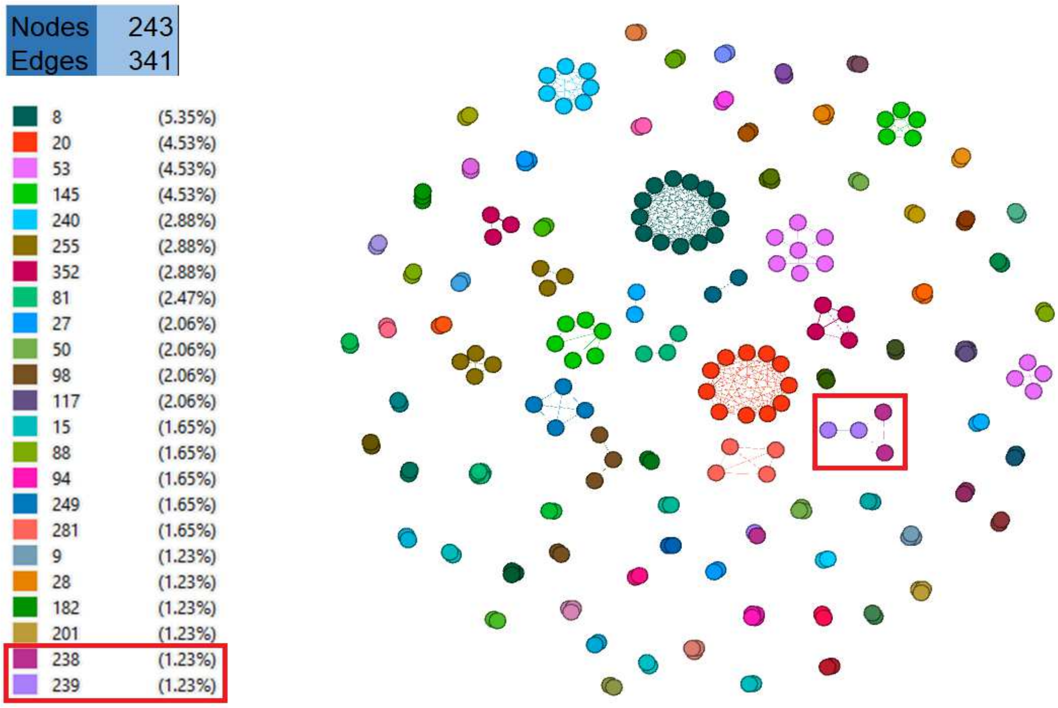


Modularity	0.9
Total Communities	817

🟦	12	(8.89%)
🟩	8	(8.37%)
🟥	4	(6.79%)
🟪	9	(4.89%)
🟧	10	(4.74%)
🟪	22	(4.38%)
🟥	13	(4.07%)
🟩	3	(3.92%)
⬜	32	(2.97%)
⬜	30	(2.19%)

With users classified according to their communities, IntelDigitalMarketing found possible relationships between users, using the graph from user communities to detect new edges between nodes considering the community. Figure 50 shows the result of the generated network of possible relationships between Twitter users. The result of the analysis highlighted possible relationships between users from the same community and from different communities, such as users from communities 238 and 239, highlighted in Figure 50. Through CS7, IntelDigitalMarketing also recommends new relationships between users of Digital Influencer Communities (FR3). As a result, these predicted data can be used to anticipate decisions and support the dissemination of information on the network.Unfortunately, in this evaluation, the prediction of relationships was not possible in the OSNs Youtube and Instagram due to the significant number of nodes used for the analysis. The computational infrastructure used for the analysis was not able to supply the memory demand. In future IntelDigitalMarketing versions, this must be considered. Therefore, Figure 50 was also generated using the Gephi tool, considering the processing costs.

Figure 50 – Graph of predictions of relationships between twitter users.



Using the CS5 query (Table 9), IntelDigitalMarketing was also able to extract the data property hashtag from the publications of the most populous community of each OSN. Table 10 presents the 5 most cited keywords in the most populous community of each OSN, queried from the ontology. On Twitter, it was the C34 community, with 1,159 users. It was observed that the most cited keyword on Twitter is “queijoazul”, from the "Artesanal Paulista" category, with ~87% citations. On Instagram, it was the C1 community, with 6,693 users. The most cited keyword was “Queijo Artesanal”, followed by “queijo” in the "Queijo" category.

We can also observe new keywords such as “instaqueijo”, “minasgerais” and “instavinho”. These new keywords were used by users in the publication using the character “#”. On Youtube, we had the C12 community as the most populous, with 14,202 users. The most cited keyword was “queijodamarisa”, which is also new.

This automatic discovery of new keywords can support marketing strategies meeting the FR1 requirement. For example, the keyword “instavinho” can be included in the marketing of Artisan Cheeses on Instagram. These new keywords can also help you discover trends. In addition, new paths to marketing dissemination can be generated according to the relationships and connections between OSN users through the collected social media (FR2).

Table 10 – The 5 most cited keywords of cheese in the most populous community of each OSN

OSN	Keyword	Category	Frequency
Twitter C34 community	queijoazul	Queijos Artesanais	~87%
	queijocoalho	Coalho	~3%
	queijocanastra	Queijos Artesanais	~2%
	queijoserro	Outros	~0.8%
	queijomofo	Outros	~0.8%
Instagram C1 community	queijoartesanal	Queijos Artesanais	~4%
	queijo	Queijos	~3%
	instaqueijo	Novas	~3%

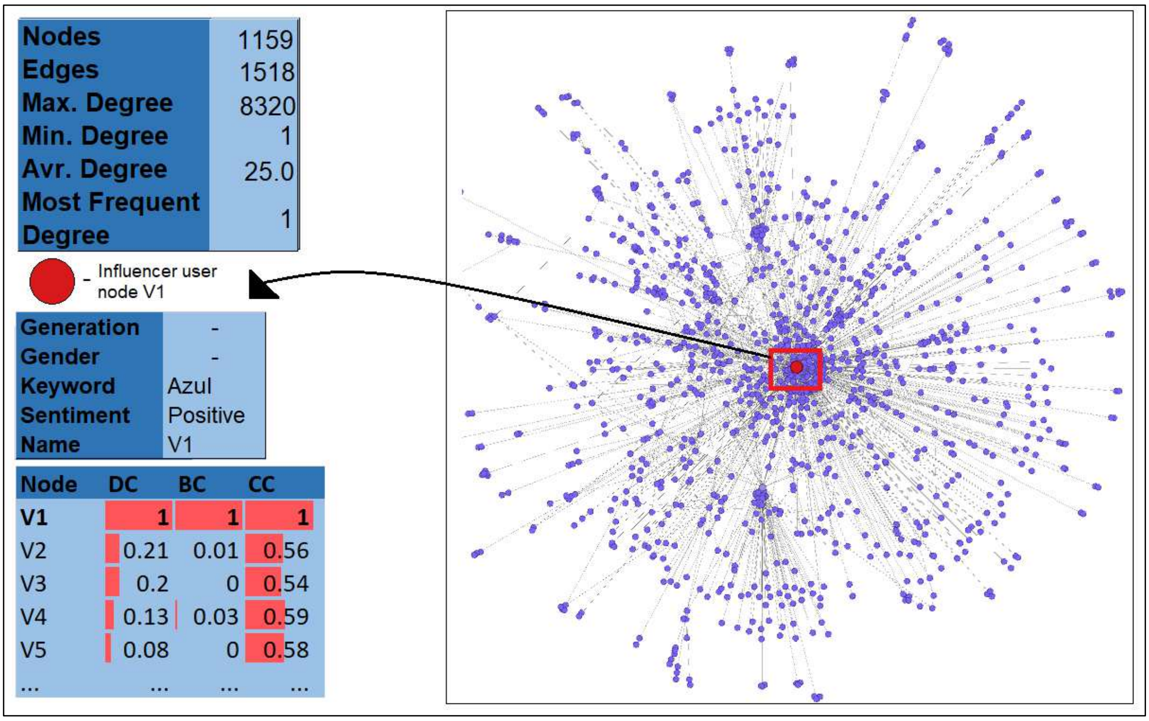
	minasgerais	Novas	~3%
	instavinho	Novas	~2%
Youtube C12 community	queijodamarisa	Novas	~84%
	gueto	Novas	~13%
	queijopingodeamor	Queijos Artesanais	~0.2%
	queijopadrevictor	Queijos Artesanais	~0.2%
	queijovendanova	Novas	~0.05%

Considering C34, C1 and C12 communities (each one the most populous of each OSN), as parameters for the CS8 query, IntelDigitalMarketing processed the influence analysis. It generated a subgraph for each community along with its general information and the ranking of influential users.

The subgraph SG34 corresponding to C34, can be seen in Figure 51, emphasizing the most influential node in red. It is a connected network, which means that all nodes have a possible path between them. The result of the influence analysis is a list with the centrality score for each node. User scores in the ranking were normalized between 0 and 1 and, in bold, node V1, which is the user that has the greatest potential for influence in the C34 community, as the values of its metrics are higher when compared to other users.

In the subgraph SG1 (Figure 52) corresponding to C1, it is also a connected network. The user with the greatest potential for influence is node V1. The SG12 subgraph (Figure 52) shows the C12 community. The minimum degree is also greater than 0 and it is a connected network, and its most influential user is V1.

Figure 51 – Twitter most populous Community's graph. IntelDigitalMarketing web interface screenshot



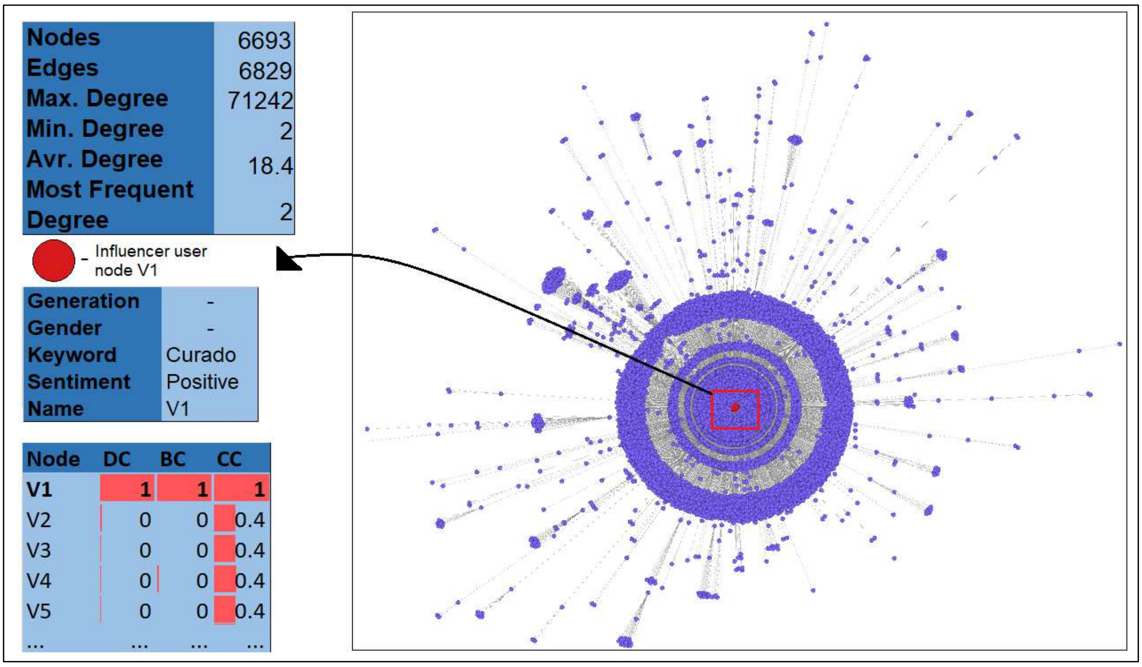
Nodes	1159
Edges	1518
Max. Degree	8320
Min. Degree	1
Avr. Degree	25.0
Most Frequent Degree	1

● - Influencer user node V1

Generation	-
Gender	-
Keyword	Azul
Sentiment	Positive
Name	V1

Node	DC	BC	CC
V1	1	1	1
V2	0.21	0.01	0.56
V3	0.2	0	0.54
V4	0.13	0.03	0.59
V5	0.08	0	0.58
...	...	...	...

Figure 52 – Instagram most populous Community's graph. IntelDigitalMarketing web interface screenshot



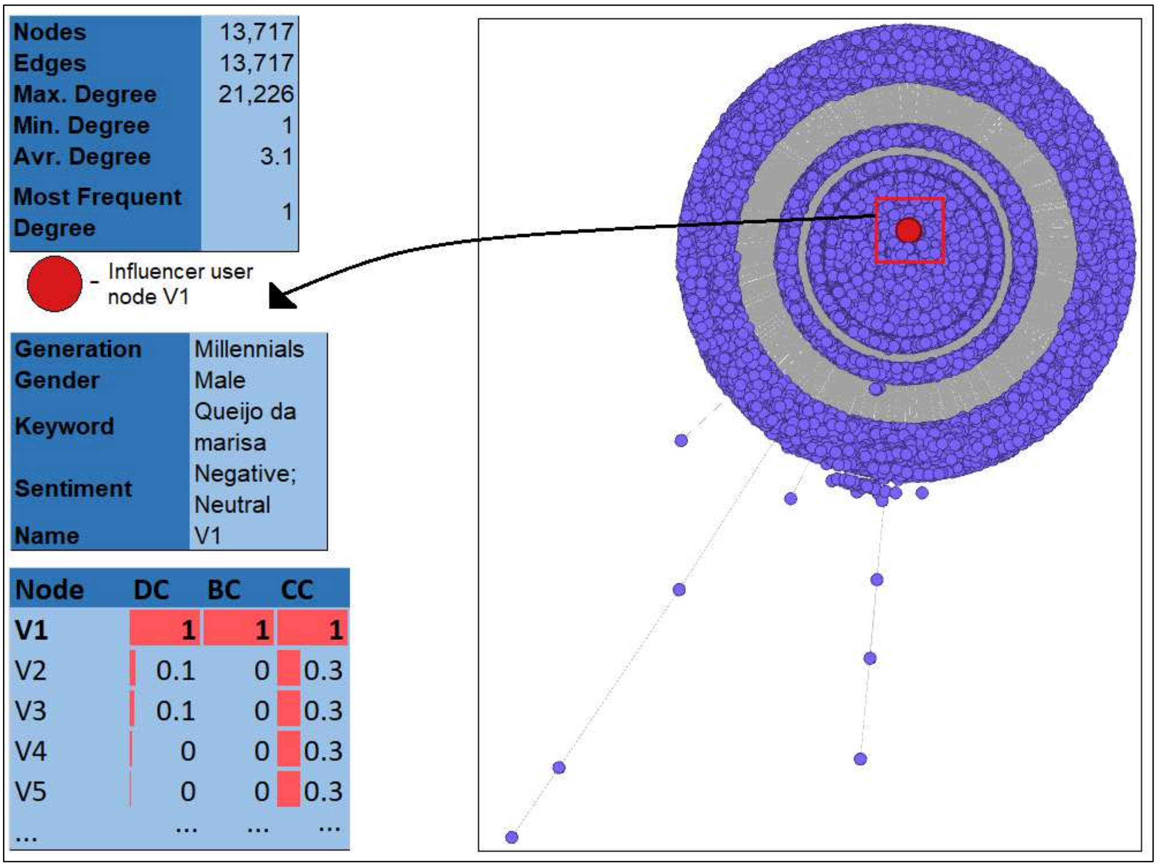
Nodes	6693
Edges	6829
Max. Degree	71242
Min. Degree	2
Avr. Degree	18.4
Most Frequent Degree	2

● - Influencer user node V1

Generation	-
Gender	-
Keyword	Curado
Sentiment	Positive
Name	V1

Node	DC	BC	CC
V1	1	1	1
V2	0	0	0.4
V3	0	0	0.4
V4	0	0	0.4
V5	0	0	0.4
...	...	...	...

Figure 53 –Youtube most populous Community's graph. IntelDigitalMarketing web interface screenshot.



Nodes	13,717
Edges	13,717
Max. Degree	21,226
Min. Degree	1
Avr. Degree	3.1
Most Frequent Degree	1

● - Influencer user node V1

Generation	Millennials
Gender	Male
Keyword	Queijo da marisa
Sentiment	Negative; Neutral
Name	V1

Node	DC	BC	CC
V1	1	1	1
V2	0.1	0	0.3
V3	0.1	0	0.3
V4	0	0	0.3
V5	0	0	0.3
...	...	...	...

5.3 RESULTS

Based on the execution of this evaluation, responses to research questions could be provided. To analyze these results and summarize our findings to answer SRQ1 and SRQ2, we answered the specific questions proposed in the evaluation.

a) Which cheese is the most cited among communities? To answer this question, IntelDigitalMarketing used community detection, topic modeling and, ontology features. From the processing of these modules in IntelDigitalMarketing, we could observe that the visualization of classified users grouped according to their interactions contributes to identifying consumers with similar characteristics in different OSNs. Through CS5, it was possible to generate Table 10 (or word clouds) for each community of each OSN, enabling the search for the most talked products. As a result, producers, and researchers could identify specific demands for products and thus intensify their production. Thus, the most cited cheeses among the communities of each OSN are: (i) “queijo azul” for Twitter; (ii) “queijo artesanal” for Instagram; and (iii) “queijo da marisa” for Youtube.

b) What is the feeling (positive or not) of users in the most populous cheese community concerning dairy derivatives? To answer this question, IntelDigitalMarketing used sentiment analysis and ontology features. During the execution of the sentiment analysis on the publications extracted from the ontology, as shown in Figures 33, 34 and 35, we could observe a positive feedback for the three analyzes carried out in the OSNs. In this way, we combined CS1 with other queries to generate specific graphs and tables, and better understand the key aspects of the products that users like or not. This analysis, used in combination with other analyses, such as community detection, allows the identification of products wanted by specific communities, which can help with strategies for future product launches and/or suspencion of production of other products. For example, we have the charts in Figure 54 that were generated for the Twitter community 34, where the predominant sentiment is positive.What is the profile of users who talk about dairy derivatives? IntelDigitalMarketing used to answer this question, the age/gender detection and ontology features. By querying the ontology, we extracted the users, considering gender and generation (CS6). From this data, it was possible to draw a user profile that talks about a specific product or the profile of users of a specific community. For example, on Twitter, we have that the majority of users are millennial men who talk mostly about “Queijo Azul” and has a positive impact as shown in Figure 54 and Figure 55.

Figure 54 – Twitter Community C34 sentiment pie chart and top three keywords bar chart presented using the IntelDigital Marketing web interface

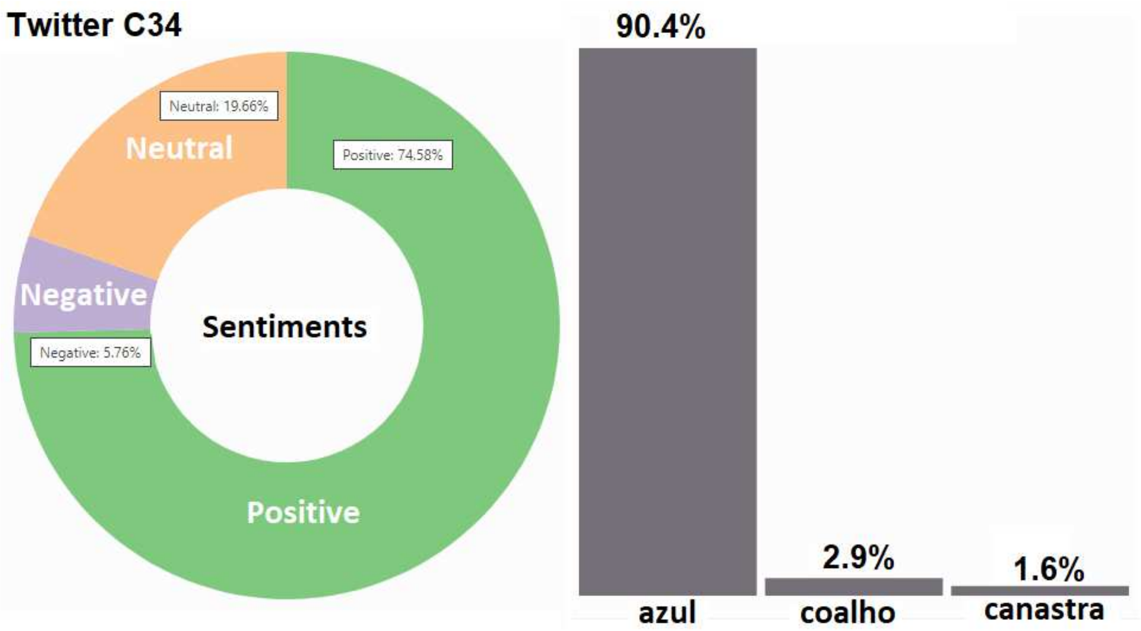
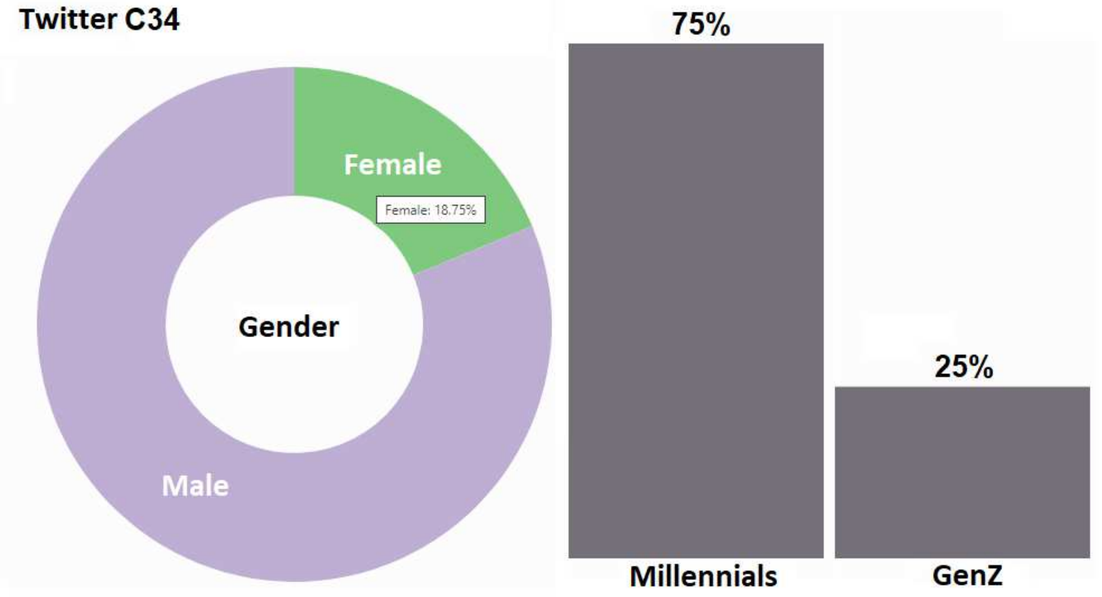


Figure 55 – Twitter Community C34 gender pie chart and generation bar chart presented using the IntelDigital Marketing web interface



c) Who are the users who can disseminate dairy derivatives in other communities? IntelDigitalMarketing used link prediction on the generated community graph to answer this question. Using the CS7 query, we extracted the object property “canSocializeWith” from the ontology to generate a network of users with significant probabilities of being related between communities, as seen in Figure 50. Then, it was possible to identify the users with the highest chances of disseminating a target product in other target communities, helping to identify potential influencers capable of developing strategies to disseminate the product in other related communities. On Twitter, relationship detection highlighted users from communities 238 and 239 highlighted in Figure 50.Who are the biggest influencers interested in dairy derivatives at a specific OSN and what they talk about? For this, we used the influence analysis. Through the CS9 query (Table 8) to the ontology, we could identify influencers in the network of users who talk about cheese on Twitter. In Figure 51, "user v1" was the user in the most populous community with the highest score considering in the metrics used. The same could be identified on Instagram (Figure 52) and YouTube (Figure 53). Thus, organic marketing strategies can be traced, considering the stratification of influencers and the products they like the most.

5.3.1 Research Questions

Considering the specific questions presented in the previous section, RQ1 and RQ2 could be answered.

SRQ1. "How is the marketing effect of the most cited artisanal cheese among the communities of the different generations of users who talk about cheese on the OSN?"
Could be analyzed, with the help of the previous answers to questions (a) "Which cheese is the most cited among communities?", (b) "What is the feeling (positive or not) of users in the most populous cheese community concerning the dairy derivatives?" and (c) "What is the profile of users who talk about dairy derivatives?". Based on the identification of the most cited cheeses in the different OSNs, from question (a), it was possible to draw a demographic profile of the different generations, considering the answers to question (c), and the effect (positive, negative, or neutral), using question (b), on the discussions in these generations-stratified communities about cheese. As a result, identify targeting digital marketing campaigns to specific groups was possible. Viewing users grouped according to their interactions and ranked content, helps researchers and producers to identify consumers and content relevant to dairy marketing by directing posts to a community, recommending a product to a community, or a combination of these. For example, a marketing advertisement for artisanal cheese in the C1 Instagram community (Table 9, Figure 45) targeted to a millennial female audience that enjoys wine. In addition, knowing whether the community's opinion of a product is positive

or negative, helps to decide whether a product is worth continuing in production or whether it should undergo changes. The CS1, CS5, and CS7 queries can be used to generate graphs or spreadsheets for each community, enabling the search for the most cited dairy derivatives by communities. Thus, the recommendations can help the producer in meeting customer demands and monitoring the market in the OSNs. Therefore, it was possible to answer RQ1, creating an overview of cheese-related digital marketing.

SRQ2. "Who are the biggest influencers of cheese at OSN? "

Considering SRQ2, we can use the influence analysis to the most populous communities of the OSNs. The subgraphs corresponding to the communities could be seen in Figures 51, 45 and 46, emphasizing the most influential node. The subgraph's general information also can be seen in these Figures. The result of the influence analysis is a list with the centrality score for each node. The Figures also show a table with some of the users scored. We can find the user with the greatest potential of influence in the community because the values of her/his metrics are higher compared to other users. Thus, RQ2 could be answered. The identification of influencers in the users' network who talk about cheese on Twitter could be found, as presented in Figures 44, 45, and 46. In these Figures, the “user v1” was the user of the most populous community with the highest score in the metrics used. Thus, researchers or producers can use this information to disseminate information to the public or improve marketing strategies using specific users. Regarding digital word of mouth marketing, this user can post cheese information or make product suggestions online. The subgraphs can be generated for the other communities, searching for each community’s most significant influencer. In the end, the recommendation of the most influential user is made through queries to the ontology, visualization of graphs and tables. Besides, v1 can be used to spread information between communities with the recommended users (object property “canSocializeWith”), such as the user_1549399313 individual (Figure 21). As a result, she/he does not need to visit several different OSN profiles to search for product information and can spread information in one community starting from another.

Therefore, answering RQ1 and RQ2, we could answer the main RQ:

RQ. "How to foster Agribusiness Products by analyzing Multiple OSNs?"

Answering RQ1 and RQ2, we could better understand the impact generated in Twitter, Youtube and Instagram of cheese products’s routine. As a result, it is possible to generate scientific knowledge and foster Agribusiness Products using the proposed solution. Therefore, we could have evidence that IntelDigitalMarketing can be used as a strategy to improve the dissemination of information of Agribusiness products. More specifically, to foster the dairy derivatives products. In addition, IntelDigitalMarketing can assist in the development of new products, identification of opportunities, improving marketing campaigns.

5.4 THREATS TO VALIDITY

It is important to analyze the reliability, especially whether the results are biased. We therefore discuss some issues that can affect the validity of the results.

Construct validity. We selected metrics to evaluate the marketing effect. However, these metrics might not be better indicators for the evaluation context. We can use additional metrics to mitigate this threat, considering different contexts. Moreover, IntelDigitalMarketing mined the main OSNs and stored the data as instances of an ontology. However, the data was limited to a specific period representing a threat. Additional evaluations need to be carried out to reduce this threat.

Internal validity: The data used was related to the dairy derivatives domain during the evaluation. The results are still preliminary, and although they indicate a positive outcome, a more detailed study is needed to present additional findings. Therefore, the data used by the IntelDigitalMarketing can pose a threat. In a broader context, other metrics and data need to be used, and, as a result, we must reassess the marketing effect. Additionally, the computational infrastructure used posed a threat. Considering other computational environments can require new evaluations.

External validity: The evaluation deals with a dataset associated with a specific domain, i.e., dairy derivatives. We need to carry out evaluations considering other domains before generalizing our results. However, it is possible to identify situations where we can obtain similar evaluation results, and the knowledge acquired can be transferred to similar real-world domains.

Reliability: We presented details of the execution of the study, but some information was probably incomplete. We have made the documentation available to ensure the evaluation reruns to mitigate this threat.

5.5 FINAL REMARKS

This chapter presented an evaluation of the IntelDigital Marketing solution, answering the proposed RQ.

6 FINAL CONSIDERATIONS

This chapter aims to present the contributions of this dissertation, in addition to its limitations and future work.

6.1 SUMMARY

This dissertation presented a solution to foster agribusiness products dissemination by extracting data from multiple OSNs. We specified a solution, called IntelDigitalMarketing, to assist producers in identifying the most sought/cited cheeses, as well as their potential consumers. For this IntelDigitalMarketing implemented data analysis techniques, favoring the understanding of the data. It used ML, ontologies and complex network to analyze data and recommend content about dairy derivatives and consumers.

This solution is innovative because it identify specific user patterns, uses implicit information, i.e., information discovered processing (a) inference algorithms, (b) structural analysis of complex networks, and (c) recommendation techniques to understand agribusiness data (dairy derivative data) extracted from OSNs.

From the literature analysis, similar works that analyzes and visualizes the OSNs information related to agribusiness were not found. Thus, IntelDigitalMarketing was developed. To collect evidence of our approach's feasibility, an evaluation was carried out using OSNs data. The results pointed to the viability of the solution.

Therefore, we first identified the problem relevance as ”agribusiness lacks solutions that allow the consumer and product online data analysis to foster products dissemination”.

We related investigated works to find more accurate and representative directions for Agribusiness marketing, exploring how the use of SNA and/or SMA can support Agribusiness. The SLM found that more primary studies on SNA and SMA in agribusiness are needed to provide a better understanding of how these technologies work. This conclusion was based on findings from selected primary studies in this SLM, as well as insights gained from analyzing online social media data. These findings support the argument that social media monitoring can complement traditional methods. The traditional market research is time-consuming, expensive, and, at times, incomplete and without representation.

IntelDigitalMarketing has proven to be a scalable architecture, capable of extracting data from multiple OSNs. In addition, it can easily be used in other application contexts by creating a folksonomy and ontology related to the context.

6.2 CONTRIBUTIONS

We can cite as main contributions of this dissertation:

- Conduction of a systematic mapping of the literature, which studied the current scenario of the literature on Agribusiness supported by the SNA and/or SMA. The study identified online social media sources, platforms, and techniques for analyzing information, evaluation metrics, and the agricultural segments covered by the primary studies;
- The development of an ontology, called Artchee-O, with the aim of representing social media related to the dairy derivatives domain, generated in OSNs;
- Specification of a set of activities through the BPM-InteldigitalMarketing process, systematizing the phases of the solution;
- Development of an architecture, called InteldigitalMarketing, capable of recommending content and influential users in a given application domain;
- The creation of dashboards that help decision-makers analyze data in a systematic way, supporting strategic decisions.
- Conduction of two feasibility studies:
    - In the first cycle, exploring how SNA and/or SMA can support agribusiness, based on Twitter data;
    - As a result of the first cycle and SLM results, the conduction of a second study, extracting data and integrating Twitter, Instagram, and YouTube (multiple OSNs) data.

6.3 LIMITATIONS

During the development of this work, there was a computational difficulty in the storage and processing of inference mechanisms. Because it is an ontology with a large volume of data, the computational cost is considerably high, which generated difficulties in the data collection activities. To deal with this problem, we limit the number of individuals that are processed by the ontology reasoner before being saved in the main ontology. Even so, it was noted that the time of the return of SPARQL queries grows with the number of individuals that are stored in the ontology. The InteldigitalMarketing solution was verified only in the Python language, so it cannot be generalized to other languages. So, it can be considered a drawback.

The "User Linkage Analysis Module" also presented computational challenges due to the large volume of data. The computational power needed for the network metrics used in the module grows with the number of nodes in the network. So, to mitigate this, we decided to apply the network metrics by communities, considerably reducing the computational cost. However, even so, the relationship prediction technique was not possible for some OSNs due to the computational limitations of the server that hosts the solution.

Finally, as a major limitation, other domains must be explored since the solution was applied only in the dairy products domain. In addition, other techniques for conducting node and media analysis should also be investigated in future versions of the InteliDigitalMarketing solution.

6.4 PUBLISHED RESEARCH

The following works resulting from this dissertation were published:

- Soares ND, Braga R, David JMN, Siqueira KB, Ströele V, Campos F. Uma Arquitetura para a Recomendação de Consumidores de Queijo Artesanal Brasileiro. In: Anais do Brazilian e-Science Workshop (BreSci). SBC; 2020. p. 113–20.
  - This work presents an initial study of a recommendation architecture, based on social media, considering the Brazilian dairy market. A technical feasibility study of this architecture was carried out generating valuable information regarding changes in the structure of the architecture.
- Soares ND, Braga R, David JMN, Siqueira KB, Nogueira TDS, Campos EW, et al. REDIC: Recommendation of Digital Influencers of Brazilian Artisanal Cheese, SBSI. ACM Proceedings Series, 2021;
  - This work presented a solution, called REDIC, to identify users of the most relevant OSNs and detect communities to support the dairy sector,

specifically artisanal cheese. The solution was used on Twitter to verify the feasibility of the proposal.

- Donato, N. ; BRAGA, REGINA MACIEL ; David, J. M. N. ; Siqueira, K. ; Stroele, V. . Data Analysis in Social Networks for Agribusiness: A Systematic Review. IEEE Access, 2023 (accepted to publication) . .Also Available from: https://doi.org/10.48550/arXiv.2208.14807
    - This work investigates works that provide solutions, based on social network analysis, that can foster agribusiness. We adopted a hybrid systematic mapping to conduct the investigation.

6.5 FUTURE WORK

As future work, it is proposed to explore other domains and other techniques to support the products disseminations. Geographic information and multimedia content can be explored to define intelligent user segmentation. It is also necessary to carry out a more comprehensive evaluation with other OSNs to verify the proposed solution's scalability. With the implementation of other OSNs, we intend to search for a connection between them using user's information as a parameter. Thus, we can disseminate information on an OSN through another. Additional metrics used in the SNA can also be studied.More up-to-date machine learning models will also be explored. As a result, we can compare model outputs and find the best option for the application and OSN.This study can also be continued through a new DSR cycle, carrying out an evaluation with producers and researchers to check the effectiveness of the proposal. Moreover, exploring extensions of the folksonomy and ontology with new perspectives on the domain. New approaches may emerge for agribusiness, and not only in the dissemination of products.

REFERENCES

1. Cheng WWH, Lam ETH, Chiu DKW. Social media as a platform in academic library marketing: A comparative study. J Acad Librariansh. 2020 Sep 1;46(5):102188.
2. Kaur R, Singh S. A comparative analysis of structural graph metrics to identify anomalies in online social networks. Comput Electr Eng. 2017 Jan 1;57:294–310.
3. Król D. On modelling social propagation phenomenon. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. 2014 [cited 2022 Sep 25];8398 LNAI(PART 2):227–36. Available from: https://link.springer.com/chapter/10.1007/978-3-319-05458-2_24
4. Soares N, Braga R, David JM, Siqueira K, Stroele V. . Data Analysis in Social Networks for Agribusiness: A Systematic Review. IEEE Access, 2023 (accepted to publication). .Also Available from: https://doi.org/10.48550/arXiv.2208.14807
5. Talamini E, Murad G, Ferreira V. Merging netchain and social network: Introducing the “social netchain” concept as an analytical framework in the agribusiness sector. African J Bus Manag [Internet]. 2010 [cited 2022 Sep 25];4(13):2981–93. Available from: http://www.academicjournals.org/AJBM
6. Soares ND, Braga R, David JMN, Siqueira KB, Nogueira TDS, Campos EW, et al. REDIC: Recommendation of Digital Influencers of Brazilian Artisanal Cheese: REDIC: Recomendação de Influenciadores Digitais do Queijo Artesanal Brasileiro. ACM Int Conf Proceeding Ser. 2021 Jun 7;
7. Riquelme F, González-Cantergiani P. Measuring user influence on Twitter: A survey. Inf Process Manag. 2016 Sep 1;52(5):949–75.
8. Soares ND, Braga R, David JMN, Siqueira KB, Ströele V, Campos F. Uma Arquitetura para a Recomendação de Consumidores de Queijo Artesanal Brasileiro. In: Anais do Brazilian e-Science Workshop (BreSci) [Internet]. SBC; 2020 [cited 2021 Nov 24]. p. 113–20. Available from: https://sol.sbc.org.br/index.php/bresci/article/view/11189
9. SIQUEIRA KB. O mercado consumidor de leite e derivados. 2020 [cited 2022 Sep 25]; Available from: http://www.infoteca.cnptia.embrapa.br/handle/doc/1110792
10. Lemos EA, Portella G, Kléber S, Cabrini J, Antônio J, Maurício B, et al. Relatório

Anual 2021 ABLV [Internet]. 20211 [cited 2022 Sep 25]. Available from: https://ablv.org.br/wp-content/uploads/2022/05/ABLV-Relatorio-Anual-2021s.pdf

11. Ribeiro Nardy D, Carvalho GR, Teixeira Da Rocha D. Mercado de Leite Fluido e Queijos no Brasil: Uma Análise de 2005 a 2016. XXIII Work. 2019 Feb 21;1–5.
12. da Silveira F, Lermen FH, Amaral FG. An overview of agriculture 4.0 development: Systematic review of descriptions, technologies, barriers, advantages, and disadvantages. Comput Electron Agric. 2021 Oct 1;189:106405.
13. Widmar N, Bir C, Wolf C, Lai J, Liu Y. #Eggs: social and online media-derived perceptions of egg-laying hen housing. Poult Sci. 2020 Nov 1;99(11):5697–706.
14. Xu G, Meng Y, Qiu X, Yu Z, Wu X. Sentiment analysis of comment texts based on BiLSTM. IEEE Access. 2019;7:51522–32.
15. Kiselev P, Kiselev B, Matsuta V, Feshchenko A, Bogdanovskaya I, Kosheleva A. Career guidance based on machine learning: social networks in professional identity construction. Procedia Comput Sci. 2020 Jan 1;169:158–63.
16. Bonchi F, Castillo C, Gionis A, Jaimes A. Social Network Analysis and Mining for Business Applications. ACM Trans Intell Syst Technol [Internet]. 2011 May 6 [cited 2022 Sep 25];2(3). Available from: https://dl.acm.org/doi/10.1145/1961189.1961194
17. Uschold M. Ontology-Driven Information Systems: Past, Present and Future. Front Artif Intell Appl [Internet]. 2008 [cited 2022 Sep 25];183(1):3–18. Available from: https://ebooks.iospress.nl/doi/10.3233/978-1-58603-923-3-3
18. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Q Manag Inf Syst. 2004;28(1):75–105.
19. Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. Proc Annu Hawaii Int Conf Syst Sci. 2013;995–1004.
20. Toprak S, Can EY, Altinsoy B, Hart J, Dogan Z, Ozcetin M. Social media video analysis methodology for sarin exposure. Forensic Sci Res [Internet]. 2022 [cited 2022 Sep 25];7(2):279–84. Available from: https://www.tandfonline.com/doi/abs/10.1080/20961790.2020.1825061
21. Rogers D, Preece A, Innes M, Spasić I. Real-Time Text Classification of User-Generated Content on Social Media: Systematic Review. IEEE Trans Comput Soc Syst. 2022 Aug 1;9(4):1154–66.

22. McAfee AP, Brynjolfsson E. Big data: the management revolution. undefined. 2012;

23. Yang X, Dong M, Chen X, Ota K. Recommender System-Based Diffusion Inferring for Open Social Networks. IEEE Trans Comput Soc Syst. 2020;7(1):24–34.

24. Ma T, Zhou J, Tang M, Tian Y, Al-Dhelaan A, Al-Rodhaan M, et al. Social Network and Tag Sources Based Augmenting Collaborative Recommender System. IEICE Trans Inf Syst. 2015 Apr 1;E98.D(4):902–10.

25. Atefeh F, Khreich W. A Survey of Techniques for Event Detection in Twitter. Comput Intell [Internet]. 2015 Feb 1 [cited 2022 Dec 10];31(1):133–64. Available from: https://dl.acm.org/doi/10.1111/coin.12017

26. Batmaz Z, Yurekli A, Bilge A, Kaleli C. A review on deep learning for recommender systems: challenges and remedies. Artif Intell Rev 2018 521 [Internet]. 2018 Aug 29 [cited 2022 Sep 25];52(1):1–37. Available from: https://link.springer.com/article/10.1007/s10462-018-9654-y

27. Aggarwal CC. An Introduction to Recommender Systems. Recomm Syst [Internet]. 2016 [cited 2022 Dec 10];1–28. Available from: https://link.springer.com/chapter/10.1007/978-3-319-29659-3_1

28. Lawton HS, McGee L. Accessibility - W3C [Internet]. www.w3.org. 2005 [cited 2022 Dec 10]. Available from: https://www.w3.org/standards/webdesign/accessibility

29. Berners-Lee T, Hendler J, Lassila O. The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.

30. Bernerss-Lee. Tim Berners-Lee: The next web | TED Talk | TED.com [Internet]. TED Talk. 2009 [cited 2022 Dec 10]. Available from: https://www.ted.com/talks/tim_berners_lee_the_next_web?language=en

31. Guarino N. Formal Ontology in Information Systems [Internet]. 1998 [cited 2022 Dec 10]. Available from: https://philpapers.org/rec/GUAFOI

32. Borst WN. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. undefined. 1997;

33. World Wide Web Consortium. SPARQL 1.1 Query Language [Internet]. World Wide Web Consortium. 2013 [cited 2022 Dec 10]. Available from: https://www.w3.org/TR/sparql11-query/

34. Saxena A, Iyengar S. Centrality Measures in Complex Networks: A Survey. 2020 Nov 14 [cited 2022 Dec 10]; Available from: https://arxiv.org/abs/2011.07190v1

35. Lu J, Chen G, Ogorzalek MJ, Trajkovic L. Theory and applications of complex networks: Advances and challenges. Proc - IEEE Int Symp Circuits Syst. 2013;2291–4.

36. Kitchenham BA. Systematic review in software engineering. 2012;1.

37. Ghermandi A, Sinclair M. Passive crowdsourcing of social media in environmental research: A systematic map. Glob Environ Chang. 2019 Mar 1;55:36–47.

38. Kitchenham BA. Systematic review in software engineering. 2012 [cited 2022 Dec 10];1. Available from: https://dl.acm.org/doi/10.1145/2372233.2372235

39. Mourão E, Pimentel JF, Murta L, Kalinowski M, Mendes E, Wohlin C. On the performance of hybrid search strategies for systematic literature reviews in software engineering. Inf Softw Technol. 2020 Jul 1;123:106294.

40. Mourao E, Kalinowski M, Murta L, Mendes E, Wohlin C. Investigating the Use of a Hybrid Search Strategy for Systematic Reviews. Int Symp Empir Softw Eng Meas. 2017 Dec 7;2017-Novem:193–8.

41. Petticrew M, Roberts H. Systematic Reviews in the Social Sciences: A Practical Guide. Syst Rev Soc Sci A Pract Guid [Internet]. 2008 Jan 11 [cited 2021 Nov 24];1–336. Available from: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470754887

42. Mahoney JA, Widmar NJO, Bir CL. #GoingtotheFair: A social media listening analysis of agricultural fairs. Transl Anim Sci. 2020 Jul 1;4(3):1–13.

43. Bernal Jurado E, Fernández Uclés D, Mozas Moral A, Medina Viruel MJ. Agri-food companies in the social media: a comparison of organic and non-organic firms. Econ Res Istraz . 2019 Jan 1;32(1):321–34.

44. Colezea M, Musat G, Pop F, Negru C, Dumitrascu A, Mocanu M. CLUeFARM: Integrated web-service platform for smart farms. Comput Electron Agric. 2018 Nov 1;154:134–54.

45. Neiva FW, David JMN, Braga R, Campos F. Towards pragmatic interoperability to support collaboration. Inf Softw Technol [Internet]. 2016 Apr 1 [cited 2021 Nov 24];72:137–50. Available from: https://dl.acm.org/doi/abs/10.1016/j.infsof.2015.12.013

46. Steinmacher I, Chaves AP, Gerosa MA. Awareness Support in Distributed

Software Development. Comput Support Coop Work [Internet]. 2013 Apr 1 [cited 2021 Nov 24];22(2–3):113–58. Available from: https://dl.acm.org/doi/abs/10.1007/s10606-012-9164-4

47. Harper L, Kalfa N, Beckers GMA, Kaefer M, Nieuwhof-Leppink AJ, Fossum M, et al. The impact of COVID-19 on research. J Pediatr Urol. 2020 Oct 1;16(5):715–6.

48. Reed M, Keech D. The ‘Hungry Gap’: Twitter, local press reporting and urban agriculture activism. Renew Agric Food Syst [Internet]. 2018 Dec 1 [cited 2021 Nov 24];33(6):558–68. Available from: https://www.cambridge.org/core/journals/renewable-agriculture-and-food-systems/article/abs/hungry-gap-twitter-local-press-reporting-and-urban-agriculture-activism/2F1289FA12CDCE6821ABE69260CCD212

49. Sabou JP, Cihelka P, Ulman M, Klimešová D. Measuring the Similarities of Twitter Hashtags for Agriculture in the Czech Language. Agris on-line Pap Econ Informatics. 2019 Dec 1;4(December):105–12.

50. Sanders CE, Mayfield-Smith KA, Lamm AJ. Exploring twitter discourse around the use of artificial intelligence to advance agricultural sustainability. Sustain. 2021 Nov 1;13(21).

51. Yigitcanlar T, Regona M, Kankanamge N, Mehmood R, D’costa J, Lindsay S, et al. Detecting Natural Hazard-Related Disaster Impacts with Social Media Analytics: The Case of Australian States and Territories. Sustain. 2022 Jan 1;14(2).

52. Pan D, Yang J, Zhou G, Kong F. The influence of COVID-19 on agricultural economy and emergency mitigation measures in China: A text mining analysis. PLoS One. 2020 Oct 1;15(10).

53. Niles MT, Rudnick J, Lubell M, Cramer L. Household and Community Social Capital Links to Smallholder Food Security. Front Sustain Food Syst. 2021 Mar 3;5.

54. Chesoli RN, Mutiso JM, Wamalwa M. Monitoring with social media: Experiences from “integrating” WhatsApp in the M&E system under sweet potato value chain. Open Agric. 2020 Jan 1;5(1):395–403.

55. McLean HE, Jaebker LM, Anderson AM, Teel TL, Bright AD, Shwiff SA, et al. Social media as a window into human-wildlife interactions and zoonotic disease risk: an examination of wild pig hunting videos on YouTube. Hum Dimens Wildl.

2021;

56. Salim JN, Trisnawarman D, Imam MC. Twitter users opinion classification of smart farming in Indonesia. IOP Conf Ser Mater Sci Eng. 2020 Jul 20;852(1).

57. Ko RKL, Lee SSG, Lee EW. Business process management (BPM) standards: A survey. Bus Process Manag J. 2009 Sep 11;15(5):744–91.

58. Yang J, Zhang XD. Predicting missing links in complex networks based on common neighbors and distance. Sci Reports 2016 61 [Internet]. 2016 Dec 1 [cited 2022 Sep 25];6(1):1–10. Available from: https://www.nature.com/articles/srep38208

59. Soundarajan S, Hopcroft J. Using community information to improve the precision of link prediction methods. WWW'12 - Proc 21st Annu Conf World Wide Web Companion. 2012;607–8.

60. Huang C, Wang J. Link Prediction Based on Weight Assignments in Complex Networks. 2022 7th Int Conf Big Data Anal ICBDA 2022. 2022;210–3.

61. Schiffer E, Hauck J. Net-Map: Collecting Social Network Data and Facilitating Network Learning through Participatory Influence Network Mapping. http://dx.doi.org/101177/1525822X10374798 [Internet]. 2010 Jul 12 [cited 2022 Sep 25];22(3):231–49. Available from: https://journals.sagepub.com/doi/10.1177/1525822X10374798

62. Scott J. Social Network Analysis. https://doi.org/101177/0038038588022001007 [Internet]. 2016 Jul 2 [cited 2022 Sep 25];22(1):109–27. Available from: https://journals.sagepub.com/doi/10.1177/0038038588022001007

63. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp [Internet]. 2008 Oct 9 [cited 2022 Sep 25];2008(10):P10008. Available from: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008

64. Freeman LC. Centrality in social networks conceptual clarification. Soc Networks. 1978 Jan 1;1(3):215–39.

65. Brandes U. A faster algorithm for betweenness centrality. J Math Sociol. 2001;25(2):163–77.

66. Serengil SI, Ozpinar A. HyperExtended LightFace: A Facial Attribute Analysis Framework. 7th Int Conf Eng Emerg Technol ICEET 2021. 2021;

67. Hoffman MD, Blei DM, Bach F. Online Learning for Latent Dirichlet Allocation. Adv Neural Inf Process Syst. 2010;23.

68. Leydesdorff L. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. J Am Soc Inf Sci Technol [Internet]. 2007 Jul 1 [cited 2022 Dec 10];58(9):1303–19. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/asi.20614

69. OWL Working Group. OWL - Semantic Web Standards [Internet]. W3C.org. 2012 [cited 2022 Dec 25]. Available from: https://www.w3.org/OWL

70. Beresford Research. Age Range by Generation | Beresford Research [Internet]. Beresford Research. 2022 [cited 2022 Dec 26]. p. 1. Available from: https://www.beresfordresearch.com/age-range-by-generation/

71. Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof MD. SWRL: A Semantic Web Rule Language Combining OWL and RuleML [Internet]. W3C Member submission 21. 2016 [cited 2022 Dec 27]. p. https://www.w3.org/Submission/SWRL/. Available from: https://www.w3.org/Submission/SWRL/

72. Sinclair J, Cardew-Hall M. The folksonomy tag cloud: When is it useful? Journal of Information Science. 2008 May 31; 34(1):15–29. Available from: https://journals.sagepub.com/doi/10.1177/0165551506078083

73. Fernández M, Gómez-Pérez A, Jurista N. METHONTOLOGY: From ontologica art towards ontological engineering workshop on Ontological Engineering. In: Spring Symposium Series. Facultad de Informática (UPM); 1997.

74. Nogueira T da S, Siqueira KB, Goliatt PVZC. Mineração de dados em rede social para avaliação de tendências de consumo do queijo artesanal no Brasil. In Juiz de Fora: Universidade Federal de Juiz de Fora (UFJF); 2022 [cited 2023 Mar 23]. p. 179–87. Available from: https://repositorio.ufjf.br/jspui/handle/ufjf/12524