Social Network Analysis and Mining (2022) 12:41
https://doi.org/10.1007/s13278-022-00870-3

REVIEW PAPER

Getting decision support from context-specific online social networks: a case study

Manuela Freire[1,2] · **Francisco Antunes**[2,3] · **João Paulo Costa**[1,2]

Received: 30 September 2021 / Revised: 5 February 2022 / Accepted: 9 February 2022

Abstract
The combination between online social networks (OSN) and decision processes provides a favorable social data analysis paradigm for efficient decision support and business-processes integration. This paper presents a framework for handling OSN's contents, providing a simpler and effective approach for information retrieval and processing. The objective is to address a decision-making problem, by using that framework to extract, process, structure and analyze the OSN's data. The decision process is not only guided by OSN data, but also by social network analysis methodology and is entirely based on the communications among social media users. Our framework combines two different, though complementary, perspectives: the analysis of the interactions among users and the semantic analysis of their discourses. In addition, it aims to bridge technology and manual-based approaches, thus enhancing the possibilities for making a better use of an OSN, using free-available software. The case study, herein, aims to estimate customers' requests, solely based on their Facebook posts, showing that the unstructured data of the web's discourse can be used to support this kind of decision processes.

Keywords Social network analysis · Decision support · Semantic extraction · Online social networks

1 Introduction

Understanding people through social network analysis (SNA) plays an important role in many real-world applications, with added value for academics and organizations. However, literature on how to engage in and utilize social networks research remains quite sparse (Monaghan et al. 2017; Salmons 2017; Isson 2018). According to Alhajj and Rokne (2018), the interest in studying human social networks has exploded, more recently, with the appearance of web-based communities.

Online social media and social networks concepts are frequently confused, and therefore, it is important to establish boundaries between them. Online social media rely on online communication platforms, built on the principles of social dynamics. In contrast, social networks concepts are much older than web and digital technologies. Social networks derive from studies conducted in the fields of sociology and sociometry, to study people's relationships. However, with the development of digital technologies, social networks became associated with online social media because its analysis targets are mainly focused on web resources, such as network structure, user's interactions and/or semantics contents. In this paper, we used the concept of online social network (OSN) to bridge social media and social networks, as we address both digital technologies and people's relationships.

The main goal of this paper is to address a decision-making process supported by a context-specific OSN, using a framework to extract, process, structure and analyze the OSN's data. Such framework incorporates two important facets, namely human interaction and network structure, by combining human capabilities, SNA and automatic data mining. In addition, it aims to bridge technology and manual-based approaches, enhancing the possibilities for

✉ Manuela Freire
maria.s.freire@fe.uc.pt

Francisco Antunes
francisco.antunes@ubi.pt

João Paulo Costa
jpaulo@fe.uc.pt

[1] CeBER, Faculty of Economics, Univ Coimbra, Av Dias da Silva 165, 3004-512 Coimbra, Portugal

[2] INESC Coimbra, Coimbra, Portugal

[3] Department of Management and Economics, Beira Interior University, Estrada do Sineiro, 6200-209 Covilhã, Portugal

Published online: 05 March 2022

common managers (especially for those less IT savvy) to make a better use of an OSN, using free-available software.

The novelty of the proposed model, as detailed in Sect. 6.5, regards the analysis of the three entities of web discourse (user, post and concept) which is adjustable to specific contexts. It structures the data of users, posts and concepts (the three entities of web discourse), allowing the creation of networks that enables a global view, not only of the social interactions among users, but also of the produced dialogs. It also combines SNA techniques to visualize networks and to calculate associated metrics, NLP and common software to support all processes, using predefined semantic patterns from other systems, when available, or using directly extracted concepts from the organizational context where it is being applied. This ensures greater reliability of the data, greater control over it and allows to control the vocabulary characteristics of each organizational context. Finally, it also integrates a graph database that allows to organize data in a graph format, making it possible to create and visualize possible sub-networks, according to the purpose of the analysis.

Although a case study, where Facebook is used to sell goods, is presented, this paper is not actually about selling on Facebook or improving sales, but rather on the evaluation and accuracy demonstration of the developed framework, when applied to a context-specific situation, as well as its usage for decision support.

Studies that incorporate SNA, as well as semantic analysis of posts, are scarce and basically remain theoretical (Power and Phillips-Wren 2012; Herring 2013; Abulaish et al. 2020), essentially by engaging in a computer-mediated communication point of view (Herring 2013), through web content (Kok and Rogers 2017) or discourse analysis (Moser et al. 2013) and text analysis, rather than in the rhetorical, argumentative, semantic and pragmatic issues of the produced text. The work of Biswas et al. (2018) proposes an unsupervised method for keyword extraction, from twitter content, combining graphical visualization and SNA metrics. The authors elect the keywords by combining the frequency, centrality, position and strength of a node's neighbors to assess its importance. Duari and Bhatnagar (2020) present a supervised method for automatic keyword extraction from a single document, by using databases of abstracts and articles, from scientific domains and news published online, to build semantic networks and keywords identification. To do so, the authors used predefined and already normalized keyword lists. Other studies, that do perform semantic analysis of OSN posts, usually adopt a manual approach for treatment and coding. Erétéo et al. (2011) combined the user's network structure and its produced content, by manually adding tags to messages, to create relationships between tags, as well as between tags and users. The study by Himelboim et al. (2013) also integrated SNA and Twitter content analysis, by manually encoding 212 tweets. Réda et al. (2017) suggested a recommendation system that identified relevant information to experts (in various areas) and manually encoded 153 tweets. They extracted and sorted, also manually, keywords from one-third of the tweets they coded (50 out of 153). Other studies for semantic network analysis used a *word cloud*, instead of SNA metrics, to visualize words' level of importance and to identify keywords in the text (Tripathy et al. 2017; Chatterjee and Trumbo 2018; Ghim et al. 2018; Kleminski and Kazienko 2018; Troisi et al. 2018).

Using the provided reporting capabilities of the different OSN platforms, to manage a social presence, is yet quite difficult, or, at least, restrictive. From a business point of view, maintaining an online social presence without adequate tools (to be used in terms of the available knowledge and budgetary restrictions) for extracting potential value-added information from different stakeholders, turning it into usable business knowledge, seems a potential waste of resources. The use of OSN data offers new possibilities for supporting daily-based activities. According to Pang and Lee (2008), *what others think* has always been an important piece of information for most people, during a decision-making process. OSN has made possible, as never before, to collect directly the opinions and experiences (both personal and professional) of a wider range of people without any *formal inquiries*, therefore allowing to change the way we look at the whole decision process, as it will be addressed in Sect. 2. Tollinen et al. (2012) argue that monitoring OSN, rather than using explicit surveys, provides more objective results on people's intentions. However, according to the literature (Isson 2018; Davis 2019; Savic et al. 2019), several issues, like the management of OSN content (that grows day by day), its heterogeneity and the effectiveness of its extraction (just to mention a few), are still open and not properly taken care of. According to Marmo (2011), the combination of SNA and web mining gives an innovative degree of detail in the analysis of OSN that can be useful for decision-making support. It provides a better structuring and understanding of the logical sequence of the produced contents of a social web discourse, as expressed by Antunes and Costa (2011).

The remainder of the paper proceeds as follows. The next section presents some key concepts about decision support within the social web context. In order to make the manuscript self-contained, some relevant topics of SNA are briefly covered in Sect. 3. In Sect. 4, we present a studies overview in OSN. The proposed framework to extract, process, structure and analyze the OSN's data is detailed in Sect. 5. This framework is applied to a business situation, as described in Sect. 6, where the obtained results are also discussed. The conclusions are presented in Sect. 7.

2 Decision support within the social web data context

The current information overload, alongside with the desire to decide in real time, can hinder the efforts to understand costumers' behavior and needs. As referred by Robinson et al. (2015), the ability to understand and analyze highly interconnected data is a key factor for companies to outperform their competitors. However, as Monaghan et al. (2017) refer, although SNA provides an increasingly detailed view of the relational and structural properties of organizational activity, the means to manage and analyze its application are still scarce. According to different authors (Akar and Dalgic 2018; Isson 2018; Davis 2019; Liu et al. 2020; Wang et al. 2020), in business, it is important, maybe more than ever, to deal with data quickly and effectively, in order to understand customers, monitor processes and support decision-making, thus increasing the need to monitor communication channels like OSN.

One of nowadays most common techniques to collect data (regarding, for instance, marketing purposes like testing new product acceptance, determining the level of client satisfaction, accessing after-sales quality, etc.) is the direct survey (whether personal or online-based). Nonetheless, it is largely recognized that this type of approach possesses intrinsic problems, as people do not always reveal their true opinions or intentions, especially in face-to-face situations (whether from politeness, lack of courage, fear of eventual consequences or simply because the interviewer is so nice or beautiful…), which may lead to significant errors in predicting activities, such as future sales or poll results.

The Latin expression in vino veritas suggests an easy way to solve the earlier problem. Although we do not advocate that we should offer a few drinks to every subject in a statistical sample just to get them *to talk*, the idea of collecting people's true opinions seems far more feasible within the context of OSN, than within face-to-face environments, as people can use made-up profiles to express their ideas, instead of using their *official* ones [please see Tollinen et al. (2012)]. Regarding the earlier stages of a decision-making process, especially at the intelligence and design stages, as defined by Simon (1977), this means getting better information quality, thus enhancing the possibility of better or more reality-tuned decisions, even though direct surveys have its own advantages, such as simplicity, cost and simpler data processing, that OSN content analysis has not. There are several limitations, that cannot be disregarded, concerning the treatment, organization and retrieval of the involved information, and therefore, the task of collecting and analyzing the content of OSN (or web discourse), commonly known as posts, remains quite challenging, because of the:

- Processing of posts' text—processing the textual data within posts requires tools to clean and standardize them, in order to capture the semantic aspects that allow to go beyond the mere identification of keywords (Robinson et al. 2015). The ideas contained in posts can be diffuse and sentences full of characters and punctuation to emphasize diverse situations. In such messages, incomplete sentences that end with ellipses ("…"), many blanks, smiles, abbreviations and slang can also be found. These characteristics strongly limit any attempt to use NLP tools (thesaurus, dictionaries, ontologies, etc.) or any computer-based linguistic analysis. As reported by Isson (2018), unstructured data analytics is a new challenge.
- Semantics—usually, posts do not share common ontologies, as they are created and changed constantly. The inexistence of standards to express web data semantics hinders the possibilities for integrating applications to analyze them. In addition, the discursive interaction within OSN, often produces information in an informal and unstructured language, that social tagging, used in folksonomies, fails to address. Moreover, users assign *bizarre* meanings to identify a situation or an object. This uncontrolled *vocabulary* used in folksonomies may suffer from ambiguity regarding the meaning of tags (Antunes et al. 2014; Freire et al. 2017).
- Posts' dimensions (number of characters)—in order to speed up communication, people tend to reduce the number of typed characters to express an idea (Freire et al. 2015), which poses a greater strain in semantic extraction. In addition, a single post may have a considerable number of associated replies (from different users), which increases the difficulty of analyzing and extracting data from OSN.
- Data characteristics—OSN is much more than a simple set of links and texts. They are interactive and dynamic complex networks, with several links that correspond not only to users and *friends* (followers), but also a set of links between posts, videos, photographs, etc. (Robinson et al. 2015; Velde et al. 2015). In addition, a user can comment posts in a successively way and then share them. This raises issues in visualization and graphical analysis. The collected data fall into three categories: structured (the users), semi-structured (the posts) and unstructured (the concepts expressed within posts). We are facing a new type of data found on OSN and this affects not only the data collection and processing, but also how to interpret these data and their analysis.
- Data analysis—the use of mathematical techniques, matrices and graphs, when analyzing OSN, allows to represent and describe networks in a compact and systematic way, thus accelerating data manipulation. In this case, SNA can provide a visual (qualitative), as well as a

mathematical (quantitative) analysis of human relationships. This movement between qualitative and quantitative methods reminds us that even when we are faced with a large volume of data, the small samples that integrate them can point to narratives that go against the *big picture*.

The decision model proposed by Simon (1977) summarizes the decision process into four phases. In each phase, this model is susceptible to the use of methods and tools from organizational and technological perspectives, where the OSN is the technological environment, and the organizational side is related to the decision-making process. The value of OSN data can be unveiled with SNA, which can provide new insights and opportunities to assist decision support and decision-making processes.

OSN has made possible the direct recollection of opinions and experiences from a wider range of people without any *formal inquiries*. Considerations on SNA usage for decision support are fragmented in the literature, not providing a complete vision of the elements that enable organizations to support decisions. To bridge this gap, we present an integrated framework to address the decision process, using SNA, as detailed in Sect. 5, in which we consider SNA for decision support, as a recurrent and iterative tool that addresses all of the four phases of the Simon (1977) decision model, namely Intelligence; Design; Choice; Implementation and Evaluation:

1. Intelligence—Data extraction can be related to the intelligence phase, while involving the environmental analysis, whether continuously or intermittently, and including activities that aim to identify situations problems and/or opportunities. During this phase, social data are collected from OSN sources for future processing. Such data can be useful for gathering information on the problem. So, the intelligence phase is the process of extraction and storage data from a specific context to recognition and classification of a problem or decision. Therefore, in this phase, SNA can improve the ability to view/explore OSN data, filtering data from a global network or from different sub-networks. Visualization techniques are also made possible by SNA, helping to translate data into something more understandable and usable.
2. Design—This phase includes problem understanding and analysis of available alternatives, based on their viability. To support this phase, SNA metrics and the analysis of multiple networks allow for a richer and more structured view of those involved in web discourse, as well as for determine relevant information. In addition, SNA metrics can be the basis for obtaining alternatives or courses of action in decision-making.
3. Choice—This is the process of choosing among alternatives designed in the previous phase. SNA and the visualization of OSN can be used for this purpose by analyzing and visualizing the concepts used in web discourse (semantic network) for usage in follow-up decisions, as it can allow efficient access to key information. Moreover, interpreting users' connections (within this phase) not only requires the analysis of the connections among users, but also their connections to posts. This measurement is extremely important as it can help to make informed decisions in several business areas.
4. Implementation and Evaluation phase—The last phase of the decision process is the implementation and evaluation of the chosen alternative. After identifying alternatives and having applied the criteria and rules to evaluate and implement them, SNA can assist decision-makers in providing follow-up information, on the network's reaction/behavior after implementation, by comparing previous and subsequent SNA metrics.

In summary, information that is produced within an OSN can be captured, stored and analyzed to support decision-making and to be made available to be used at any time. SNA can play an important role in business context, by structuring information and creating/revealing alternatives.

3 Using social network analysis

SNA contemplates several metrics for studying OSN. According to Fu et al. (2017), data mining refers to the automatic discovery of interesting structure in data and it includes techniques that intend to infer, from data, models that meet specific objectives. Accordingly, and because social data are involved, it is important to understand how to mine it, in order to retrieve useful information (Samanthula and Jiang 2014), to support organizational or personal decision-making. As referred by Batagelj et al. (2014), temporal network data must be expressive in terms of substance and social context, implying that it has to be carefully selected in order to be relevant. As expressed by Wasserman and Faust (1994), there are properties and methods that are only associated with users, which include how they stand out within a group, quantified through metrics such as centrality and prestige, their level of expansiveness and their popularity parameters.

According to Erétéo et al. (2011), Wasserman and Faust (1994), SNA metrics are generally appropriated for several levels of analysis and result's interpretation can be subdivided according to the object of study. However, in web discourse analysis, it is important to divide it into structural and semantic, thus delimiting the context where the discursive interaction takes place. By collecting information regarding

Table 1 SNA metrics and properties—managerial cross-reference

SNA metrics	Interpretation of Centrality measures	Activity	Control information flow	Importance	Influence	Power	Prestige (Status, Popularity)	Proximity
Degree	Number of links of a node. Nodes with lots of neighbors are central	●	●	●	●	●	●	
In-degree	Nodes that receive highest number of interactions in the network	●	●	●	●		●	
Out-degree	Nodes that send highest number of interactions in the network	●	●		●	●		
Closeness	Distance from a node to all others. How close a node is from all other nodes	●			●		●	●
Betweenness	Number of times a node acts as a bridge along the shortest path between two other nodes		●		●			
Eigenvector	Nodes connected to central nodes are central themselves			●			●	●
Page rank	Node is important according to the quantity and quality of links pointing to it			●	●		●	●
Modularity class	Identify subgroups or clusters, while indicating the structure density			●			●	●

existing links, it is possible to build a good image of the users' entire network, as well as their message exchange. In each case, a set of techniques should be selected according to the study objectives, as well as the topological characteristics and dynamics of the network to be analyzed.

Although SNA contemplates several metrics associated with the study of OSN, there are no right or wrong ways/indicators to approach OSN, as mentioned by Hanneman and Riddle (2005). We can utilize or combine SNA metrics to exploit OSN to support decision-making, as it can provide a deeper understanding on how network entities (its users, their posts and embedded concepts) interact all together, as well as to evidence properties of OSN interactions, using SNA metrics. Such metrics allow to gain insights on activity, control information flow, importance, power, prestige (status, popularity) and proximity within a network, as we summarize in Table 1 (Wasserman and Faust 1994; Page et al. 1999; Arif 2015; Saint-Charles and Mongeau 2018; Ruas et al. 2019; Ishfaq et al. 2021):

- Activity can be measured in incoming and outbound links, quantifying received/sent interactions by a specific node. A node with a high activity level means that it holds a high number of interactions. In contrast, nodes with low activity are peripheral in the network and they are not active since they are isolated.
- Control information flow represents a user's ability to control the diffusion of information in the OSN. This information diffusion can provide recommendations, for instance, to guide other customers to new products or options that might interest them, thereby increasing the possibilities of additional sales.
- The importance of a node is related to the ability to establish relationships with others, and it depends on the importance of its neighbors. For example, a user is important or influential if it is followed by other important or influential users.
- Power identifies who is widely involved in relationships with others and who is more visible within the network. Nodes that have a greater number of links tend to be more powerful because they can affect directly more nodes.
- Prestige is associated with the level of expansiveness and popularity of a user. A user with a high prestige level has direct contact with many others. In contrast, users with low prestige are isolated. Popularity is related to the relationships created with others (followers) and with the relationships that are created between them.
- Proximity describes how close a node is to another and how fast a user can reach everyone in the network, with the lowest number of nodes in-between. Keeping track of this type of users is an essential part of a business process, when reaching customers quickly is a premise.

4 Studies overview in social media network

In order to identify differences between frameworks for OSN analysis, found in literature, we examined several studies published over the last ten years. The studies show the increasing attention that OSN data analysis is receiving for supporting decision-making in organizational context. The review comprised selected papers presented in Table 2. Some of them present theoretical frameworks, but mostly are exploratory and standardized approaches. We analyzed them using five characteristics: OSN platform, used Software/Platforms (SW/Platforms), Entity (scope of analysis), Process Steps and SNA usage.

For the OSN platform data source, we characterized frameworks based on the OSN or type of social media from which the data were extracted. In terms of software/platforms, we characterized frameworks based on the type of tools used to extract, process and analyze data. Furthermore, we also identified the scope of analysis in terms of type of analyzed entities (user, post and concept). In other words, we have identified whether or not they have analyzed user's interactions, as well as the semantic content, though of them did not identify the type of entities. In addition, we characterized frameworks based on the process steps taken to perform all phases of the OSN analysis. Finally, we identified if SNA techniques were used or not. At a high level, this included determining whether the frameworks used SNA metrics.

The advantage of our framework, compared to those found in the literature, is that it can handle structured, semi-structured and unstructured data. Only a few of the reviewed frameworks incorporate these three types of data into a single environment, as they are usually analyzed separately. Differently, our framework is based not only on relational aspects (or user interaction), but also on data extraction, data processing and specific data analysis of the social environment in which social data are unstructured and informal. This grasps the basic social structural features that are key to a comprehensive analysis of OSN and its produced content.

The analyzed studies have used SW/Platforms that require a high level of expertise in information systems and the use of external sources, such as ontologies, for semantic analysis. In contrast, our framework uses user-friendly and low-cost SW/Platforms that can be used and easily manipulated in business context. External sources, for the semantic analysis, are dependent on already defined and structured data, based on the principle that they can be used by people and processed by machines. However, we are still far from that reality because it seems yet impossible to update and create standards (ontologies) for all kinds of existing business contexts.

The presented studies represent various researches in diverse areas, using a wide variety of methods. As the papers were mainly theoretical, they did not consider all elements, entities (scope of analysis), process steps and SNA techniques. However, in this small sample, we see the exploratory nature of OSN research. So, we concluded that there is a current need for an integrated framework that can easily analyze and summarize the huge amount of available OSN data, by using common tools.

Nowadays, the process for analyzing data from OSN is simplified, as the costs to store and process information have been reduced, hardware and software have improved, user-friendly and data analytics tools have emerged (Isson 2018). In addition, the technological development has increased the skills of people in information systems, as well as decision-makers and managers that have become more technology-savvy. In their roles, they simply need knowledge in the field of information systems and enough quantitative skills to be able to analyze and interpret data in an effective way.

5 Framework for OSN analysis

In our framework, the semantic analysis feeds itself, i.e., it uses the semantics found in texts (posts) according to each context of analysis. In addition, a semantic processing algorithm places all the transformations in one location, evidencing the applied logic. Our framework also supports a cleaning database, as well as the ability to update it with new concepts (found in each context) in intermediate and recursive steps. Moreover, the framework establishes a sequence of steps for processing and standardizing post contents, which are integrated in a semantic dimension that includes keyword identification and context analysis (differently from the other frameworks).

Compared to the few studies that have measured SNA properties, our study/report not only presents a theoretical approach, but also presents a case study where the framework is applied. With our framework, it is possible to *join* all analyzed entities (user, post and concept) into a unique network environment or separately analyze each sub-network and calculate SNA metrics. Using the framework, we can see the *big picture*, i.e., we know who interacts with whom (user to user relationship), who initiated a discursive exchange, who commented and who said what. All of this in a single network, with the possibility to analyze it with different levels of granularity.

Our framework for analyzing context-specific OSN is depicted in Fig. 1. It starts with a data extraction process, over a context-specific OSN, following three iterative stages, that generates data, represented in several networks. Executing this sequence in alignment with organizational context needs, requires simple and expedite tools. For instance, in

Table 2 Type of frameworks found in literature and their characteristics

Author	OSN Platform	SW/Platforms	Entity (scope of analysis)	Process steps	SNA usage
Gjoka et al. (2011)	Facebook	HTML scraping	User	Data Extraction, Data Processing, Analysis	No
Zielinski et al. (2013)	Twitter	Twitter Streaming API, MySQL, Apache QPID, Weka, NLTK toolkit	User, Tweet	Data Extraction, Data Processing, Analysis	Yes
Oussalah et al. (2013)	Twitter	Twitter Streaming API, MySQL, Apache Lucene, WordNet, PostGIS, Django, Python	User, Tweet, Semantic content	Data Extraction, Data Processing, Analysis	No
Aladwani (2014)	Facebook	–	User, post	Data Extraction, Data Processing, Analysis	No
Vosecky et al. (2014)	Twitter	URL link, Twitter REST API	User, Tweet, Semantic content	Data Extraction, Data Processing, Analysis	No
Alkhyeli and Mansour (2015)	Twitter, survey questionnaire	–	–	–	No
Banica et al. (2015)	Twitter	Apache: Flume, Sqoop and Hadoop, NoSQL, Gephi, NodeXL	User	Data Extraction, Data Processing, Analysis	Yes
Caroleo et al. (2015)	Twitter, Facebook	Cassandra, NoSQL databases, MongoDB	–	Data Extraction, Data Processing, Analysis	No
Fernando et al. (2015)	–	Social network APIs	Semantic content	Data Extraction, Data Processing, Analysis	No
Lai and To (2015)	Webpages	WordSmith, Lexicmancer, SPSS	Semantic content	Data Extraction, Data Processing, Analysis	No
Madan and Chopra (2015)	Facebook	JVM (Java)	User	Data Extraction	No
Sarker et al. (2015)	Survey of the literature	–	–	Data Extraction, Data Processing, Analysis	No
Sathick and Venkat (2015)	BSAU	R-tool, SQL	User, Semantic content	Data Extraction, Data Processing, Analysis	No
Ghafoor and Niazi (2016)	Questionnaire	–	User	Data Extraction	Yes
Ma and Che (2016)	–	JVM (Java), Xpath (XML)	User, Semantic content	Data Extraction, Data Processing, Analysis	No
Vicario et al. (2017)	Facebook	Facebook Graph API, IBM WatsonTM AlchemyLanguage service API	User, Pages	Data Extraction, Data Processing, Analysis	No
Walha et al. (2017)	Twitter	Thomson Reuters Open Calais, Dbpedia	User, Tweet, Semantic content	Data Extraction, Data Processing, Analysis	No
Appel et al. (2018)	Claims database	–	User (physicians, patients), health providers	Data Processing, Analysis	Yes
Ruas et al. (2019)	Facebook	ELKI, SPSS, Gephi, NodeXL	User, Post	Data Extraction, Data Processing, Analysis	Yes
Mahanti et al. (2012), Chan et al. (2020)	Pirate Bay, YouTube, Hulu; Warez-BB	Excel, Python	User, Post	Data Extraction, Data Processing, Analysis	No
Arafeh et al. (2021)	Twitter	Neo4j	User, Tweet, Semantic content	Data Extraction, Data Processing, Analysis	No
Adikari et al. (2021)	Twitter	Python, Word2Vec, GSOM	User, Tweet, Semantic content	Data Extraction, Data Processing, Analysis	No

Table 2 (continued)

Author	OSN Platform	SW/Platforms	Entity (scope of analysis)	Process steps	SNA usage
Alhalabi et al. (2021)	Twitter	R statistical software, Twitter4j java API, Stanford parser, Weka API, NodeXL, Gipher	User, Tweet, Semantic content	Data Extraction, Data Processing, Analysis	No
Tabassum et al. (2021)	Twitter, Facebook, Instagram, Google	Algorithms Space Saving and Biased Random Sampling	Semantic content	Data Extraction, Analysis	Yes

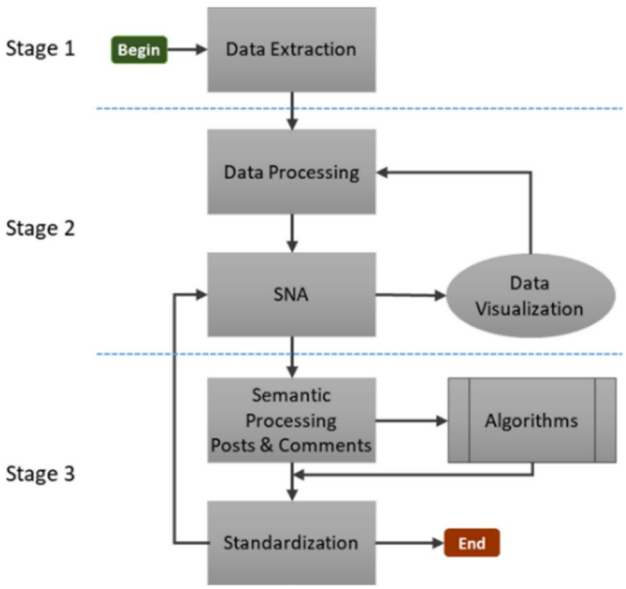


Fig. 1 Framework workflow (Freire et al. 2017)

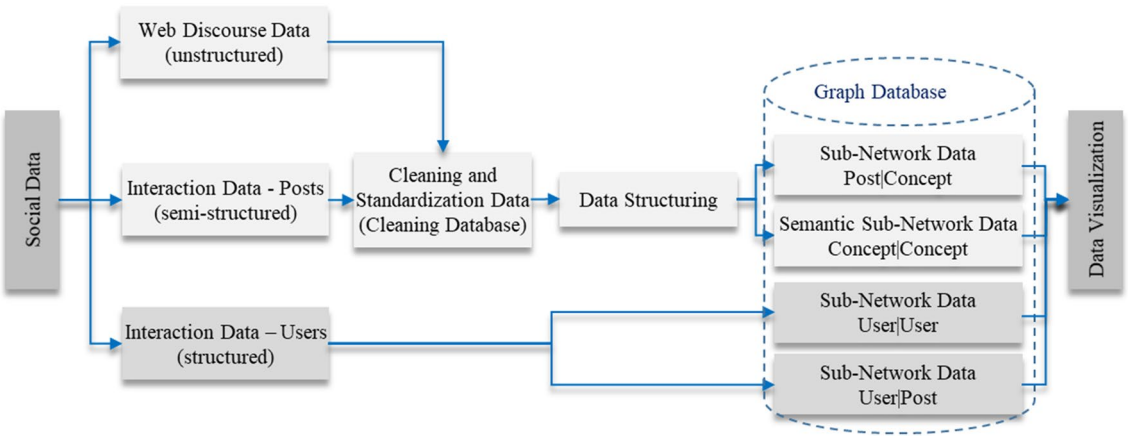


Fig. 2 Workflow to structure social data

stage 1, a variety of tools available online or the Graph API of the OSN can be used to extract data. This tool allows to extract tabular files of users' activities from the posts of the OSN. Stage 2 uses a spreadsheet (we used MS Excel) to store and manipulate data, as well as to identify and standardize network concepts in stage 3. The algorithm used in stage 3 uses Excel VBA. In stage 2, Gephi[1] is used, not only for data processing, visualization and manipulation of networks, but also for calculating the most important SNA metrics.

Figure 2 represents a conceptual overview of the data structuring procedure incorporated into the framework. It consists of a set of recurrent and iterative steps of framework's stages 2 and 3. It is also illustrated that Stage 2 involves not only *Data Processing,* but also the representation of *Social Data*. During this step, the structured data (typically *user Interaction Data*) are stored directly in the *Graph Database*. In contrast, unstructured *Web Discourse Data* and semi-structured *Interaction Data* (typically posts) are cleaned and followed by a *Standardization* process in Stage 3, which is previous to network graph analysis and the final execution of *Data Visualization* techniques. The main reason for all these steps is to structure data according

[1] https://gephi.org/.

Table 3 Online social network data extraction tools

Programming knowledge	Tools	Link	Characteristics
	Digitalfootprints	http://digitalfootprints.dk/footprints.dk/	Restricted access with login and license
	Discovertext	https://discovertext.com/	
Not Required	Infoextractor	http://www.infoextractor.org/	Open source
	Facebook Graph API	https://developers.facebook.com/docs/graph-api/overview	Open source
	NodeXL	https://www.smrfoundation.org/nodexl/	Shareware
	Nvivo/Ncapture	https://www.qsrinternational.com/nvivo/home	Open source
	Facebook Python SDK	http://facebook-sdk.readthedocs.io	Open source
	Facepager	https://github.com/strohne/Facepager	Open source
Required	Pattern	http://www.clips.ua.ac.be/pattern	Open source
	Rfacebook	https://CRAN.R-project.org/package=Rfacebook	Open source
	SocialMediaMineR	https://CRAN.R-project.org/package=SocialMediaMineR	Open source

to their type and to prepare it to be exported into network visualization tools.

5.1 First stage—data extraction

The first step is to obtain all interactions among users (user|user), their own posts (user|post), as well as all data regarding posts and comments. Such data make it possible to study the social interactions (user|user), as well as the semantics of posts. To this purpose, there is a large variety of tools (even free of charge), which can extract a wide range of elements for each user account, in distinct formats (namely, social data regarding what people do—instead of what they say they do—and information deliberately made available by them, for example, when they post a message). The available data within a OSN detail (sometimes meticulously) users' activities. Some of the available tools to extract such information are evidenced in Table 3.

The processing of unstructured data identifies and removes redundant data, using a cleaning procedure to enhance the final standardization of data. Afterward, data processing allows to structure all data and to store them into a final database that encompasses the necessary information and structure to represent a network, by using a network visualization software. Such database, known as a graph database, as defined by Harrison (2015), Kemper (2015), allows leveraging complex and dynamic relationships between data, thus creating knowledge and competitive advantages (Robinson et al. 2015).

5.2 Second stage: making sense of data

5.2.1 Data processing

The preliminary processing of OSN data intends to structure the previously stored data. It is necessary to transform the preliminary data into:

- Structured data, which refer to users' interactions.
- Semi-structured data, which refer to users and their posting activity.
- Unstructured data, which regard to concepts contained within the posts.

It is known that networks represent systems where entities (nodes) are interconnected through links (Wasserman and Faust 1994; Savic et al. 2019) and that most networks are defined as one-mode networks, in which all nodes belong to the same set. However, when a network is composed of several entities (distinct sets or levels), as the ones that need to be extracted, such networks are referred as a two-mode networks, also known as an affiliation or bipartite networks (Ikematsu and Murata 2013; Opsahl 2013; Banerjee et al. 2017). As networks of this kind can represent two or more entities that belong to distinct sets, they are used to analyze the three levels of interactions among network entities, namely the interactions of: users; users and posts; as well as posts and concepts. The latter are text characteristics extracted from documents, whether manually or using preprocessing routines, which are described in Sect. 5.3.2, in order to identify single words, multi-word expressions, whole sentences or syntactic units, which are categorized by specific identifiers, as defined by Isson (2018).

For semantic analysis, user-generated content (captured as pure text) is used to guarantee the distinction between

textual content, relevant for the analysis, from other forms of communication, namely images, and/or icon formats (smiles, emoticons, etc.) that require specific processing. For example, image data would require specific software to understand image content and translate it to textual content (see Bouet et al. 2009; Ignatov et al. 2017). Therefore, different kinds of data require different processing algorithms, or different analytical methods, as used in Yu et al. (2019) and Adib et al. (2021). The obtained data allow a posterior study of the social interaction between users, as well as of the semantics regarding the contents of their posts. The main goal of this stage is to associate a post with a user and, in turn, a concept with a post, to determine who said what.

For a simultaneous analysis of the three entities/levels of the network (the interaction between: users; users and posts; as well as, posts and concepts) a transformation of the two-mode network into a one-mode network is required (Opsahl 2013; Banerjee et al. 2017). This is done by firstly building a two-mode network based on the relationships between the users and posts and, secondly, by building another two-mode network with the entities based on the relationships between the posts and the concepts found within them and, finally, by joining both two-mode networks, transforming them into a one-mode network. This can be represented by a rectangular matrix M_{ij}:

$$M_{ij} = \begin{cases} 1 \text{ if there is a connection between } a_i \text{ and } a_j, \ i \neq j \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

where $a_i, a_j \in N = N_1 \cup N_2 \cup N_3$ and N_1, N_2, N_3 are the sets of users, posts and concepts, respectively.

After the previous process, nodes with null in-degree and out-degree (representing irrelevant data from unconnected nodes, usually derived from automatic messages) are eliminated. Afterward, it is necessary to join the remaining data, namely all interacting users and respective posts, with the posts and respective post replies, to create a single network. At this stage, SNA helps to gain useful information about the network structure and relationship patterns, such as who talks to whom, what content was transmitted and who is connected to whom. Golbeck (2015) states that SNA software provides both visual and mathematical analysis of human relationships, contained within a network, and the communication exchange embedded within textual discourse.

6 Data visualization

According to Pang and Lee (2008), the summary of documents in natural language can be verbal and/or visual. Therefore, we can consider that a graphical representation of a network, encompassing all users, posts and concepts, is a summary of the produced discourse. The visualization of those elements allows to represent the information in different layouts, as well as to obtain summaries, highlighting the most representative concepts within posts (by using SNA metrics), according to the corresponding semantic network data. Data visualization simplifies the process of grasping and interpreting such information. In addition, it allows using the results in exploratory tests and, thus, analyze and adjust data as needed, letting a more intuitive data interpretation of the distinct levels of analysis, helping to accelerate not only data processing, but also the semantic analysis described in the next subsection.

To facilitate visual evaluation and demonstration, Tabassum et al. (2021), fixed the size of samples in 1000 edges (though they also referred dynamic sampling techniques). Other authors (Fu et al. 2017; Gaeta 2018) addressed this challenge by capturing snapshots, to explore the relation between entities (users and posts). Therefore, different software products can help to understand large complex networks, containing hundreds or even thousands of nodes and edges. Such tools not only foster the graphic representation and analysis of large-scale datasets, in a technical way when needed, but also contributes to other situations that just might require a quick visualization to understand a specific dataset. Most network visualization tools do not require programming skills, while producing a high-quality visualization, even when handling large datasets. The challenge of visualizing social networks data depends not so much on its size, but rather on the computational capacities of the computer (particularly CPU and RAM).

By default, SNA analysis software has network visualization modules, which perform data exploration using various layouts, colors, sizes and other properties that are assigned to both nodes and links. For instance, modularity class, as previously stated, allows to identify subgroups or clusters, while indicating the structure density; closeness centrality, reflected in nodes' size, indicates differences and independence in participation and allows to show the interaction flow (using edges' thickness) in which users communicate with others, through a minimum number of intermediaries; centrality degree sheds light upon the group's network structure and the number of exchanged posts between users; in-degree and out-degree metrics allow to differentiate received links (in-degree) from exit links (out-degree), in the presence of oriented networks, and clean irrelevant posts; the betweenness centrality contributes to verify the information flow and the cohesiveness of the network and whether nodes are central, or even indispensable, to the network.

Using visualization tools for data analysis requires an interactive recurrent process, where human involvement is essential for a systematic analysis of the obtained results, as well as to verify if adjustments are necessary. The visual analysis is used both to eliminate irrelevant data from the users' network and to identify important nodes in the

semantic network. This is done by using distinct and/or interconnected graphs, which encompass users and semantic networks, allowing to perceive the summary of the produced discourse between users, as well as associating each one with their posts. The semantic summary, constituted by a sub-network, shows an overview of the original post and it explores the OSN structure and semantics contained in web discourse. The nodes of both networks and sub-networks are discriminated by assigning different weights to identify and select the most relevant elements of the semantic network.

6.1 Third stage: semantic analysis

One of the key challenges of processing unstructured data from OSNs is the unavailability of standard concepts. Although there are standards available in online datasets (such as WordNet, NLTK Corpora, etc.), they are not able to grasp all formal language, thus presenting limitations when being applied in specific contexts. Traditional techniques and algorithms existing in NLP [i.e., latent semantic analysis, hashtags, dictionary-based methods, lexical chain analysis, statistical approaches difference-in-difference, topic extraction, etc. (Liu et al. 2010; Sakaki et al. 2010; Bapna et al. 2018; Chua et al. 2019)] are difficult (or even impossible) to apply in context-specific situations, because they rely on complex algorithms [i.e., *K*-means, hard mo-VMF, Kalman filter and support vector machine algorithm (Sakaki et al. 2010; Rosa et al. 2011)], or they have a large dependence on external lexical sources [i.e., WordNet, Word2Vec, Wikipedia, NLTK library, Apache Lucene index, SenseClusters (Banerjee et al. 2007; Liu et al. 2010; Navigli and Lapata 2010)], to interpret semantic data. To address these challenges, the developed framework uses a cleaning database and an algorithm to perform the semantic analysis.

In contexts as different as a company that sells strawberries or another as an airline company, the used language is completely different and, therefore, defined and normalized standards (ontologies) within the semantic web that could be used by all are hardly attainable. Therefore, while the semantic web is not fully implemented, it is necessary to create cleaning databases, which constitutes a comprehensible database to support NLP (Provost and Fawcett 2013), according to each context, to respond in an expeditious way to the reality of each business context. For this reason, this process is not automatic and depends on someone who really knows the network's context to incorporate into the process the used language, so that it can be applied automatically.

We need human intervention because, as Wachsmuth (2015) pointed out, the assembly of algorithms depends on the information to be found, which is often ad hoc known. The knowledge (language) of the context-specific is captured manually and transferred from the specialist (human) to the *machine*. This knowledge is stored in the cleaning database to be used automatically by the *machine* until the entire process of cleaning and standardizing the data is concluded. The process of updating cleaning databases is a time-consuming one and it bears a steep learning curve, especially when OSN content is used, to ensure algorithms are ready to be employed.

In the next subsection, we describe the algorithms created for data cleaning purposes and semantics processing.

6.1.1 Algorithm for data cleaning

Extracting concepts from real-world data is a challenging task, and it implies the creation of a cleaning database. The cleaning database encompasses concepts that depend on the business context. The cleaning database is created to process posts' unstructured data. It feeds the cleaning and standardization algorithms and it is used to: interpret and process the text; extract the semantic network from the web discourse, created by the interaction between users; obtain a semantic network to perceive the connection between concepts within the discursive exchanges; and clean and standardize text (e.g., by removing spaces and eliminating unnecessary words). Therefore, the cleaning database contains tables of stopwords, smiles, synonyms and punctuation. It should be noted that the use of standards (ontologies) or a cleaning database must be defined and fed for each decision problem and each decision context.

In short, the algorithm allows filtering and refining data so that it becomes usable and searchable. After cleaning textual data, it is possible to create semantic networks with summaries of the discursive exchanges and networks of concepts, which fits into a broader analytical framework for decision support. The algorithm encompasses the following steps:

1. Initialization: the algorithm allows to select the input area where posts are and the output area where the processed posts will be stored.
2. The data cleaning process can be stated as follows:
    a. A cycle that removes any non-alphanumeric character, punctuation or various space between concepts (dots, commas, semicolons, dashes, parenthesis, etc.).
    b. A subroutine that searches for *smiles* in posts and replaces them with a standardization term defined in the cleaning database.
    c. A subroutine that searches for punctuation characters in each post, replacing them with a standardization term defined in the cleaning database. During this phase, the algorithm validates the amount of sequential punctuation characters that it finds, to standardize them according to defined cases. Punc-

tuation usually marks the limit of sentences but, as far as web discourse is concerned, it can be used to emphasize what is being written and could represent, therefore, exceptions in the use of language.

d. A subroutine that searches for stopwords in posts and removes them according to the cleaning database. This subroutine is very important, as there are frequently used words that should be filtered as part of the text cleanup process (e.g., function words such as *am*, *the*, *a*, and *of*).
e. A subroutine that normalizes synonyms in posts, according to a synonym table, because when the meaning of two different words are identical or nearly identical, it is possible to substitute the concepts with a single one. The synonym table is also used to replace possibly confusing concepts using a more standardized form (e.g. Online Social Network and OSN) and to represent variations of the same concept, which may exist regarding a unified key concept.

Finally, the output is created with previously defined data headers, as well as extracted concepts from posts and their identifiers (IDs). Throughout the data cleaning process, a matrix is built, where rows correspond to the extracted concepts and columns to the associated attributes with each concept, namely a sequential ID, regarding the order in which the concept is found in posts; a post ID, to associate the concept with the post where it appears; the concept itself; and a new ID that identifies whether it is a concept from a post.

6.1.2 Standardization

Text standardization is one of the steps in text preprocessing, normally a complex task in discourse synthesis (Lukanin 2015; Isson 2018). After identifying irrelevant data, the cleaning database is configured to discard them in a second processing. The remaining data are stored in a graph database for analysis and manipulation which will be explained in the next paragraphs.

The next step, keyword detection, is important for identifying most common and relevant concepts in posts (Aggarwal 2011). For that purpose, a new network needs to be created, encompassing users, received posts and the concepts within. In such network, each concept becomes an entity with its own ID. Concepts with the same semantic meaning are then sought (knowing that each concept has a different ID) and the same ID was assigned to all concepts bearing the same meaning, by replicating the links that associated a concept to a post. At this step, graphical visualization of the network is found to be very helpful.

A network of concepts is then created, encompassing users, posts and concepts within. Every concept in this network is unique, allowing to count (using the out-degree metric) the number of times each concept is posted or linked. In addition, variables k_i ($i = 1, \ldots, n$) are considered as the numerical value associated with each of the concepts, allowing the global result to be calculated from the product between variable k_i and the out-degree metric for the concept i.

6.1.3 Semantics processing

The first step, at this stage, is to transform posts into networks of words, as a basis to perform a semantic analysis, using text mining and SNA. Text mining creates structured data from unstructured text (Wachsmuth 2015; Isson 2018). A simple way to transform text into SNA interpretable data is to split every post into multiple pieces by defining each word as a node of a semantic network.

To interconnect concepts to posts, two steps are needed. The first one obtains a summary of each discursive exchange, while the second one identifies the network of concepts (this process is done resorting to relational algebra). To summarize the content of posts (first step), it is necessary to identify neighboring words. The result of this process is a network, in which every concept is a node, regardless of its existence in another post. By doing so, the content of each post is summarized through a semantic sub-network, constructed by establishing a relationship between the pairs of words encompassed in each post, as represented in Fig. 3a. The text within each post is mapped like a word chain. The first word of the sentence links to the second, the second to the third, and so on, i.e., the chain becomes a list of nodes providing that every node has an edge to the next node. To detect concepts (second step) and to look for specific information, another network, as the one depicted in Fig. 3b, is created. In this network, each concept is an entity with its own ID. For this approach, each concept should be a univocal node, as it only occurs once within the entire network. As all concepts with the same semantic meaning now have the same unique identifier, it is possible to link them to the post or posts where they appear.

While in the first step each post is associated with a semantic sub-network, in the second, any post can connect with the complete semantic sub-network, as shown in Fig. 3b. The integration of both structured (the network of users) and unstructured data (the textual data) into a single analysis environment allows to obtain a more refined level of detail, regarding, for instance, information such as customer consumption trends. This means that businesses will be able to react and make faster decisions regarding a product or customer. In addition, this approach lets us see the *big picture*: it is possible to know who interacts with whom (through the links between users), who initiated the discourse (by following the links between users and posts),

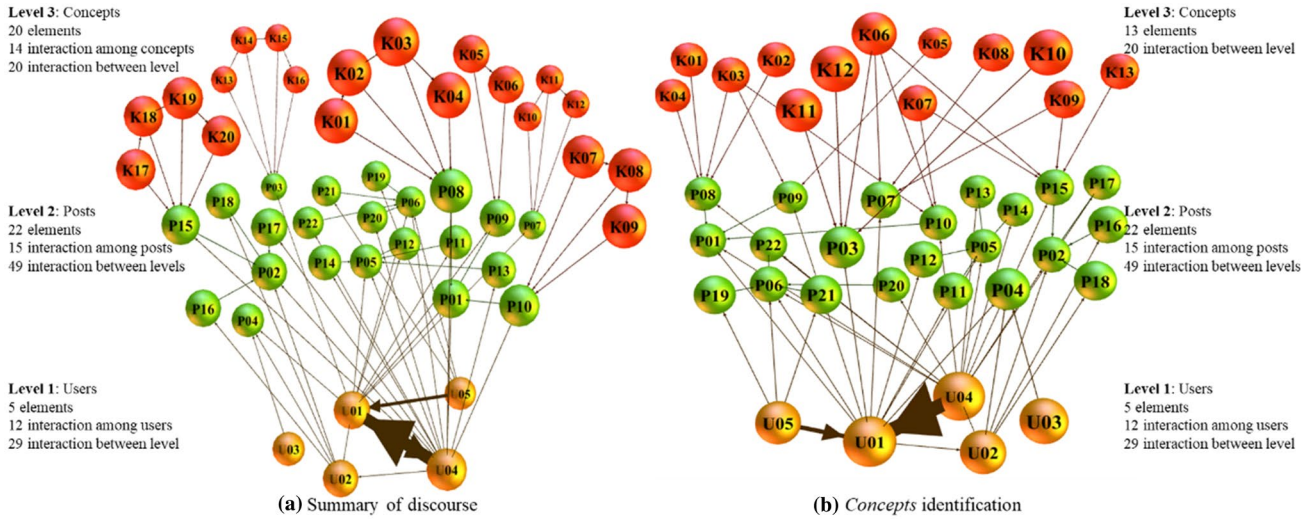


Fig. 3 Two-mode Network with 3 entities (user, post and concept)

who commented on what and who said what. All of this, in a single network, can provide organized and structured information, about people (for instance, customers' and sellers' interactions), namely for decision support.

7 Case study analysis: a strawberries sale

To apply the previously described framework for supporting business decisions, we chose a simple idea, namely a sale of freshly picked strawberries. To implement it, a strawberry producer was contacted and made ready to deliver strawberries at customers' workplaces. A box of strawberries, with an attached and appealing message, was delivered at potential customers' workplaces: "Directly from the producer to the customer. If you want some, join the group created on Facebook and leave your request". As those potential customers (users) joined the group and asked for strawberries, the associated quantities were counted and recorded for subsequent data validation. Requests were accepted and delivered on the same day. Please notice that costumers' requests do not encompass any type of order form, but the simple informal use of Facebook posts and that, upon delivery, the costumer could change its mind and actually buy a lesser quantity or none at all, or even a greater quantity (depending of the seller's availability, at the moment of delivery).

However simple the idea of selling strawberries may seem, it encompasses several management issues (request validation; management of requests modification; product availability checking; delivery logistics; customer satisfaction; etc.) that can affect business performance and profitability. Although being a very simplified view of a business problem, it does yet provide enough ground to perform a semantic analysis of the discursive exchanges and to verify if it could provide a correct recommendation for the required quantities to transport and deliver. The implicit simplicity originated lesser discursive exchanges within the group and shorter posts. Accordingly, fewer events were created in Facebook, narrowing the number of elements to analyze.

7.1 First stage: data extraction

Data were gathered using the Facebook Graph API Netvizz that were publicly available (no longer publicly accessible since September 4th, 2019). With this API, we extracted tabular files regarding users' posts over the created Facebook group: a two-mode network that showed posts, users and connections between them; a network with all interactions between users (a one-mode network); and a tabular file where its rows contained the text posted by users.

In accordance with Facebook settings, restrictions, privacy policies and API specifications, we only used online available data of the users within the created group; no information that bore any privacy restrictions, defined by the user, was included nor made available in the dataset. The page from which data were extracted was a restricted Facebook group, exclusively accessed by its members, following an administrator's validation. The contents that were produced within the group were also restricted and only made available to registered members, unless users' privacy settings specified otherwise. In this case, the content may have been available to others. At this point, we want to stress out that user data was anonymized for ethical and legal reasons (Chester et al. 2018; Tripathy and Baktha 2018), in order to guarantee users' privacy and data usage in agreement with national and international laws.

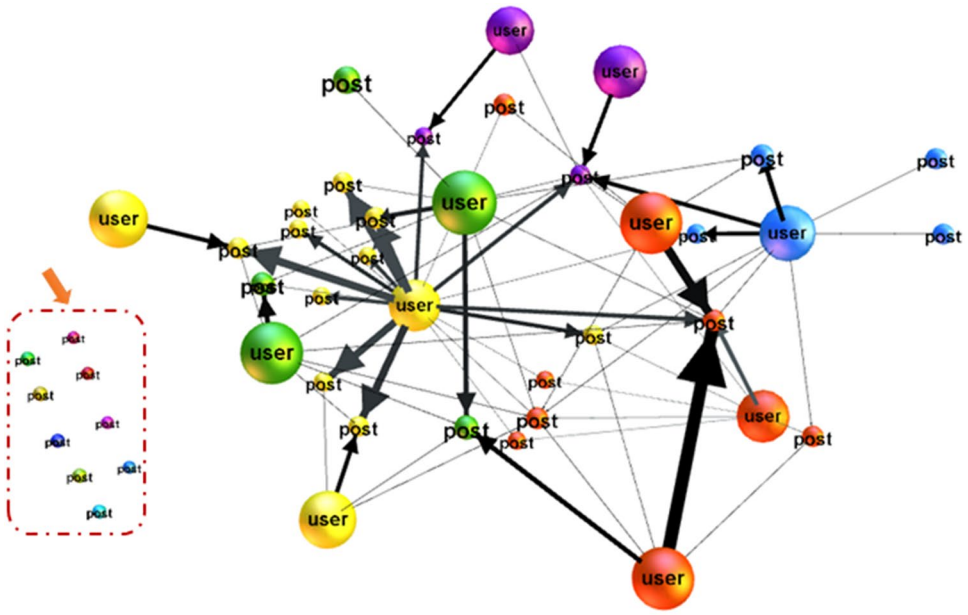


Fig. 4 Network 1—users and their posts

With the extracted data, we created two different networks, stored in a simple Microsoft Excel spreadsheet for later analysis using Gephi:

- Network 1—which encompassed all interacting users and their posts. This network has 46 nodes and 71 edges, as depicted in Fig. 4, encompassing all nodes, regardless of any interaction between them.
- Network 2—with all users' interactions. This network comprised of 13 nodes and 48 edges, only had users.

7.2 Second stage: data processing and interpretation

Facebook contents were captured as pure text in order to guarantee the distinction between textual content, from images, smileys, emojis, etc., which required preprocessing. For ethical, legal and privacy issues, data were anonymized by replacing user names. The identification and standardization of network concepts were performed using Microsoft Excel VBA (Visual Basic for Applications)-based algorithms.

Initially, the data from Network 1 and Network 2 were imported into Gephi without any processing, to identify irrelevant posts. Gephi allowed to create real-time visualizations to explore the network and instantly see the results. The software also allowed to apply automatic classifications over the network's visual appearance, in terms of colors and size of entities (users and posts).

The posts of Network 1 (Fig. 4) that had no connections to any other post or user were deemed irrelevant and eliminated, as they derived from Facebook itself. After deleting those posts, data were reanalyzed using Gephi. Network 3, which derived from Network 1, was formed by all users, who had at least one connection to a post, as depicted in Fig. 5. This two-mode network had 36 nodes (11 users and 25 posts) and 67 edges. Modularity class was used to identify existing subgroups (Savic et al. 2019) and five distinct communities were identified. The seller itself constituted a single class, while users with the same color belonged to a community, as easily identifiable in Fig. 5 (in this case, it was possible to determine that they worked at the same place). Closeness centrality allowed to identify participation differences. To evidence the amount of interactions between entities, the in-degree and out-degree metrics were also used.

As mentioned above, this network encompassed users who had a link with a post, regardless of any interaction between users. Considering this specificity, we had to rank users according to their out-degree (users who to respond to others). Table 4 shows the obtained results for both out-degree and modularity class metrics for all users through network 3. From the point of view of consumers, the most influential and active user is User-01 because he had an out-degree of 10, while user User-03 was the one with the lowest out-degree value, as he merely wrote one post.

The identification of the users with the highest out-degree value was very important to the seller, as he used this information to send posts to this highest-users and use the influence that they have in the other subgroups. In particular, the seller used discursive exchanges to affect positively attitudes and/or purchasing intentions of other customers. In fact, customers are more *aware* than we might assume, and the revenue impact of improved sales can increase (or in the opposite direction, decrease). Thus, out-degree can help to improve sales through modeling the OSN which results from interaction between users during social activities. Although, in

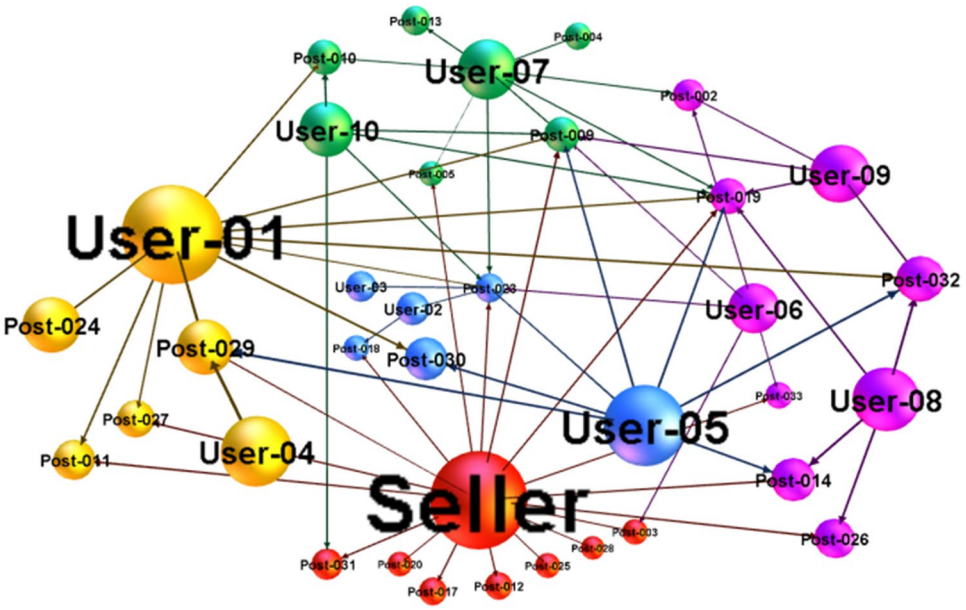


Fig. 5 Network 3—users and their posts (derived from Network 1 after initial data processing)

Table 4 Out-degree values and identified communities

User	Community	Out-degree
Seller	0	19
User-01	1	10
User-04		1
User-07	2	8
User-10		5
User-05	3	7
User-02		2
User-03		1
User-06	4	6
User-08		4
User-09		4

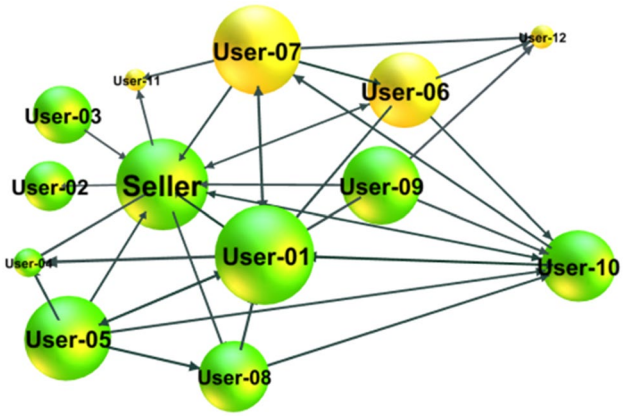


Fig. 6 Network 4—interactions user with user (derived from Network 2 with processed data)

small networks, the seller may address each of the customers, with the growth of the network, the time (cost) spent in contacting (*sting*) all customers become very large. Reaching these key customers allows an efficient dissemination of information without the need for direct action by the vendor.

Network 4, another one-mode network, as shown in Fig. 6, derived from Network 2 after data processing, included only the users that had at least one connection to another element, comprising 13 nodes and 48 edges. This network allowed to assess the relationships among users, by identifying the most significant interactions properties, namely in terms of importance, power, prestige, influence and information diffusion of users within the network, using centrality metrics.

The network properties, summarized in Table 5, allowed to conclude that the most active users were User-01, User-07 and User-05 because they had a higher out-degree (all with a value of 6). This meant that they created links between 6 other users. The out-degree is usually a measure of how influential the actor can be inside the network, which substantiates the activity that the user had to respond to others. In this business context, this is important because customers (view as users) with a high out-degree centrality are more likely to influence others and have higher power in OSN. This type of users can diffuse an opinion about the product through OSN over time among members of the group. This diffusion of information can provide recommendations and guide to other customers about new products or options that might interest them, thereby increasing the possibilities of additional sales.

Meanwhile, User-01 was the user with the highest value in the in-degree metrics (namely 8), which meant that his/her posts were viewed or commented the most. The fact that

Table 5 SNA metrics results from network 4

	Degree centrality	In-degree centrality	Out-degree centrality	Closeness centrality	Betweenness centrality	PageRank	Eigenvector centrality
Seller	19	9	10	0.91	0.33	0.12	0.98
User-01	14	8	6	0.67	0.08	0.10	1.00
User-07	11	5	6	0.67	0.07	0.08	0.81
User-10	11	7	4	0.59	0.03	0.09	0.84
User-05	9	3	6	0.63	0.00	0.05	0.51
User-06	7	2	5	0.67	0.02	0.04	0.37
User-08	6	3	3	0.45	0.00	0.05	0.38
User-04	5	4	1	0.00	0.00	0.32	0.64
User-02	4	2	2	0.50	0.00	0.05	0.26
User-09	4	0	4	0.61	0.00	0.02	0.00
User-12	3	3	0	0.00	0.00	0.04	0.24
User-11	2	2	0	0.00	0.00	0.04	0.37
User-03	1	0	1	0.50	0.00	0.02	0.00

these posts were extensively commented or displayed might ascertain their significance, what clearly shows the importance of this user. Therefore, this SNA metrics could provide relevant information for decision support, by revealing some level of *power*. User-01 had one of the highest values in the metrics closeness centrality, betweenness centrality and eigenvector centrality. The highest closeness centrality represents its independence and that s/he communicated with others through a minimum number of intermediaries. Users that have more links tend to be more powerful because they can affect directly other users. Concerning the betweenness, this user could be designated as a *leader*, as it influenced the amount of strawberries that others ordered. This fact is important for web discourse analysis as it indicates that this user could influence others. He also has great influence over what flows in the network. When this user shares information, such as opinion about the *strawberries*, s/he is influencing others. Results from the eigenvector centrality confirmed that User-01 was one of the most important users of the network. Therefore, even if a user is only connected to a few users of the network, having a low degree centrality, those neighbors may be relevant and, consequently, the user is also important when it has a high eigenvector centrality. In the opposite way, we identify User-03 as a user with low activity, with low importance, power or prestige. This user had the lowest values in almost all SNA metrics, what clearly shows that he is less important.

After analyzing the two networks separately, we combined them with the data of the tabular file (as referred previously this file contained the text posted by users) to obtain a summary of the discursive exchanges and to identify *who said what* and a new network with all entities was created. This network was an earlier two-mode network that was transformed into a one-mode network. It was constituted by three distinct entities: users, posts and concepts, as defined in the framework. With such transformation, we used SNA to analyze the three networks to reveal different aspects of the relationships between distinct entities. To extract the concepts of the posts, the cleaning database and the algorithms created for this purpose were used. The cleaning database fed the cleaning and standardization algorithms, by using tables that aided the processing of unstructured data: to interpret and process the text, to extract a semantic network for perceiving the relationships between contained concepts, and to clean and standardize the text.

The task of identifying irrelevant data was essential for a systematic analysis of the obtained results, as well as to verify if some adjustments were necessary before advancing to the next step, namely the reconfiguration of the cleaning database with more stopwords. Due to space constraints, it is not possible to represent here all created and analyzed networks, in order to completely refine data. For this reason, we only present one of them, Fig. 7.

Initially, all data were entered in Gephi, without any processing, just to identify irrelevant concepts. This *raw data analysis* is important, to know as much as possible, about the contents of the discursive exchanges. In the created network it was difficult to examine all nodes and understand the collected discursive exchanges. This network was dense and presented different topologies. As it can be seen from the Table 6, before the cleaning database optimization, with raw data, it consisted of 673 nodes and 1.148 edges (links). In a second step (network with intermediate data), and after cleaning some irrelevant data, the network had 309 nodes and 480 links. In the last step (network with final data), after cleaning all irrelevant data, the network had 237 nodes and 342 links. Initially, there were 574 concepts that, after eliminating the irrelevant data, were down to 171. 70% of the concepts were rejected because they were considered irrelevant.

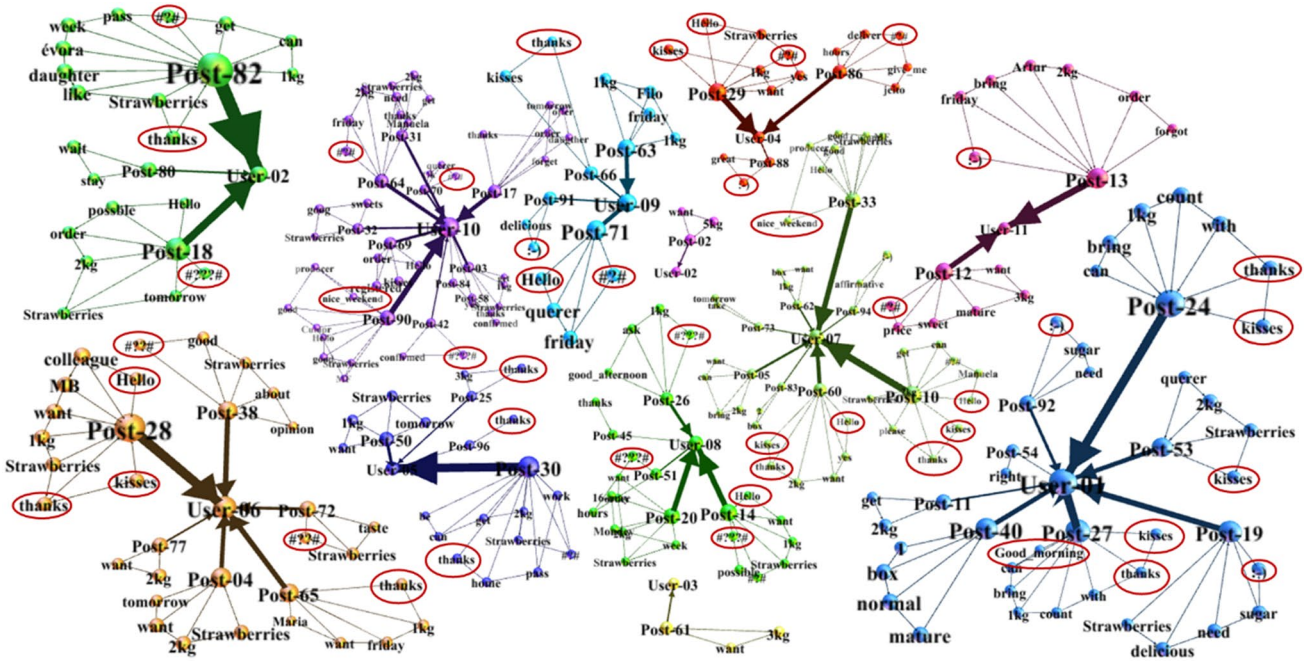


Fig. 7 Data visualization to support text mining: Network 5 with intermediate text mining

Table 6 Type and number of nodes and edges in each network

	Entities	(1) Raw Data	(2) Intermediate Data	(3) Final Data	% (3) ÷ (1)
(a) Type and number of Nodes	User	13	13	13	100%
	Post	86	56	53	62%
	Concept	574	240	171	30%
	Total Nodes	673	309	237	35%
(b) Type and number of Edges	Post|User	86	56	53	62%
	Concept|Post	574	240	171	30%
	Concept|Concept	488	184	118	24%
	Total Link by type of Entities	1148	480	342	30%

By visualizing Network 5, depicted in Fig. 7, it was possible to identify and discard irrelevant data (such as typical greetings as *hello*, *good morning*, *thanks*, etc.), allowing the reconfiguration of the cleaning database in a later process. The irrelevant data allowed the reconfiguration of the cleaning database, by discarding it, in a later process. Real social data are often large scale and considerably affected by *noise* with hidden, unobserved, or invalid nodes and links. Such noise hinders analysis by increasing the amount of data to process and hiding *important* information. So, data visualization helps to filter out noise (links, stopwords) and reveal data patterns.

Finally, Network 6, as shown in Fig. 8, was created with all the users, posts and concepts. Figure 9 depicts four examples (A, B, C and D) of semantic sub-networks, taken from Network 6, summarizing request manifestations. This final network (7) consisted of 237 nodes and 342 edges.

7.3 Third stage: semantic analysis

In this study, the identification and standardization of network concepts were performed using algorithms, which were implemented with Microsoft Excel VBA. Although the approach is relatively simple, it requires programming, NLP and text mining knowledge.

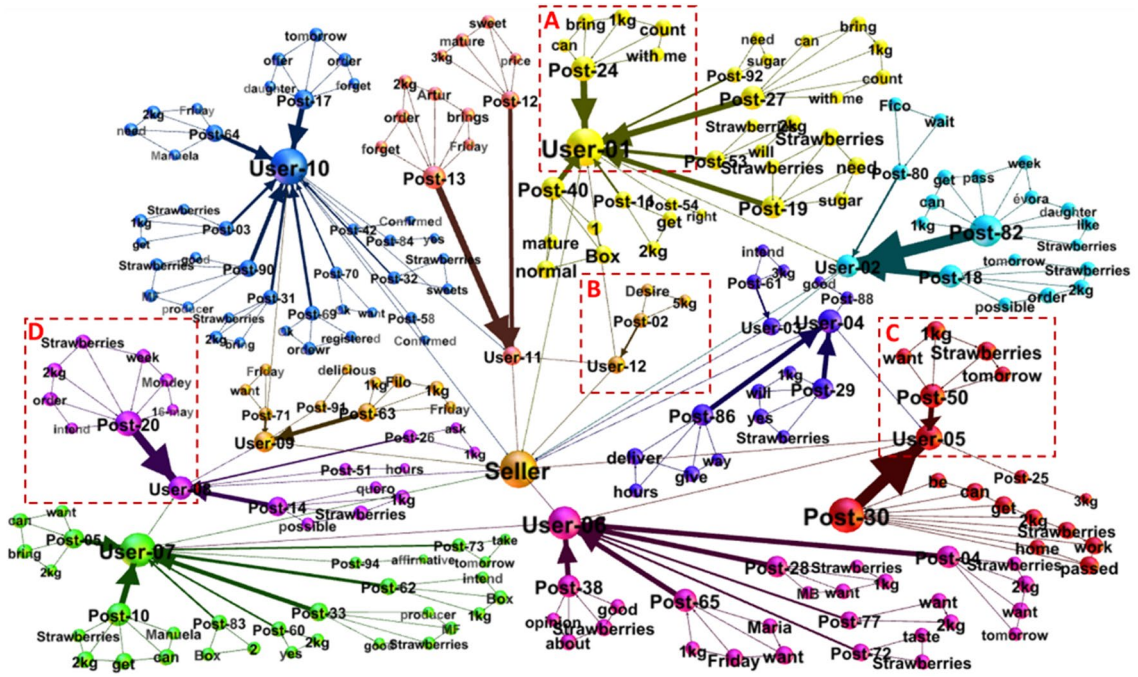


Fig. 8 Summary of all request manifestations

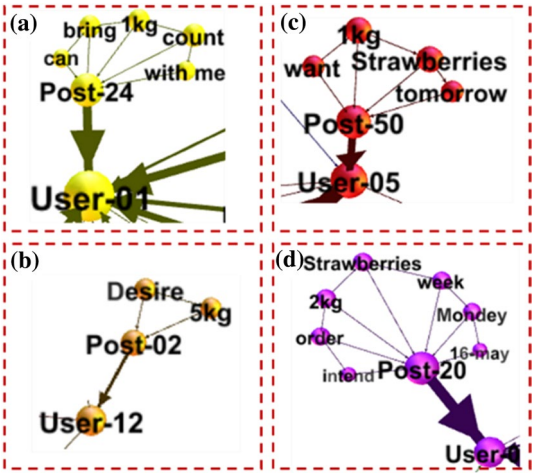


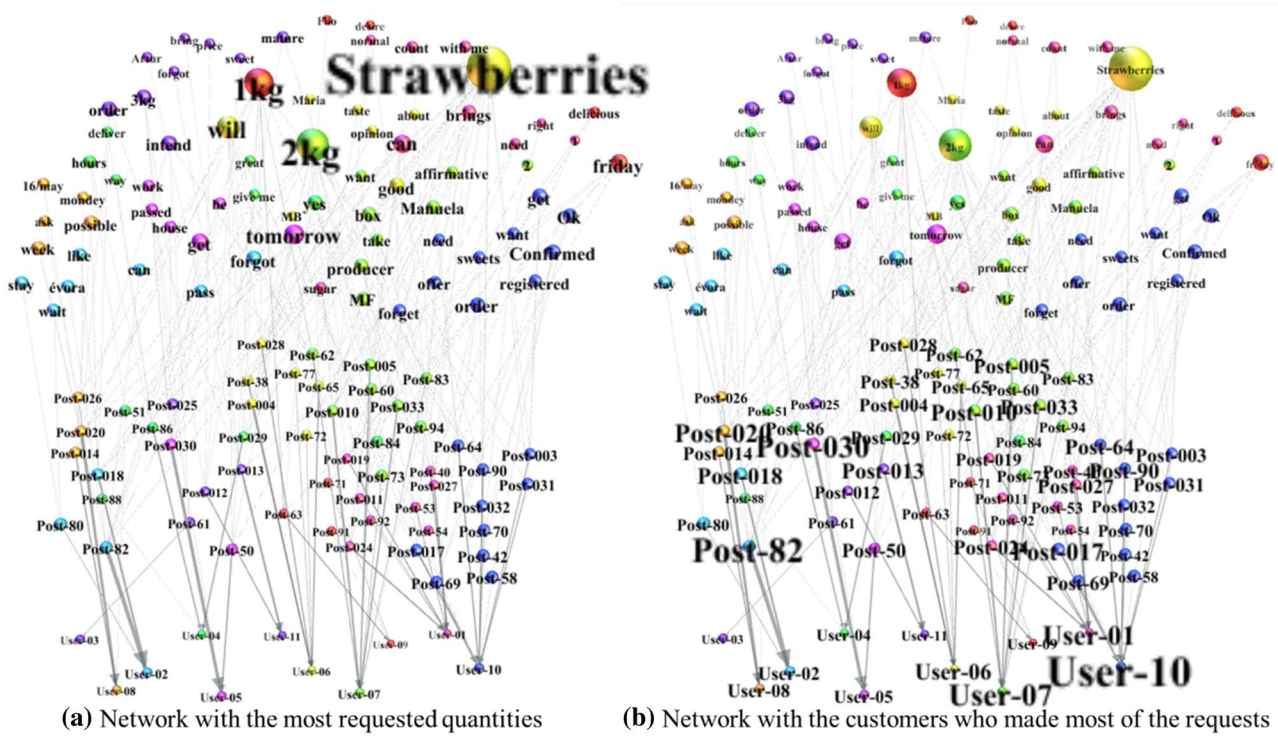
Fig. 9 Example of 4 posts with request manifestations

After processing the data, we can create a variety of sub-networks, depending on the decision problem we are dealing with. In this case, administrator's posts were disregarded leaving only costumers' posts. We wanted to extract information about customers to resume their needs and identify the requested quantities.

The process of individually extracting and analyzing concepts can provide the summary of an opinion expressed in web discourse (Pang and Lee 2008; Aggarwal 2011). Hence, with this goal, Network 7, represented in Fig. 10a and in Fig. 10b, was created, encompassing users, posts and concepts. The figures' network is the same but with different metrics. In Fig. 10a, the node and text sizes correspond to the out-degree metrics. In this network, it is possible to identify the quantities that were most requested. To identify the customers who made most of the requests, text size was defined accordingly to the in-degree metric, producing the network depicted in Fig. 10b. These networks were built with concept's unicity to identify the requested quantities and the customers who requested them.

So, to calculate the requested quantities, the out-degree metrics were used, because we were interested in the nodes that had the most direct votes, as they informed us on how many times a concept had been used. In addition, variables k_i ($i = 1, \ldots, n$) were deemed as the quantities for each of the i concepts. Moreover, the quantities were calculated from the product between variable k_i and the out-degree metrics, obtaining the results expressed in Table 7a.

To calculate the quantities requested by each customer, an attribute that identifies the relationship between *concept|post|user* was used. This relationship allowed to calculate the user's out-degree for the concept and, therefore, to identify who wrote the concept, as well as its related post. The quantities per customer are the product between the occurrences for each user of the i concept and the variable k_i, with the results shown in Table 8a.



(a) Network with the most requested quantities **(b)** Network with the customers who made most of the requests

Fig. 10 Summary of web discourse

Table 7 Quantities of strawberries requested by type of packaging

Concept (k)	(a) Framework results		(b) Manual results	Δ
	Out-degree	Kg	Kg	(a)—(b)
1 kg	12	12	10	2
2 kg	13	26	26	0
3 kg	3	9	9	0
5 kg	1	5	5	0
Total	29	52	50	2

7.4 Confronting estimated requests and actual sales

We did some manual processing in order to verify the obtained results, Tables 7b and 8b. We validated the quantities of strawberries ordered by type of packaging and the quantities of strawberries ordered by customer. We performed all calculations and logical operations manually to test the reliability of the results obtained through our framework. Note that this manual data processing is slow and error-prone, when facing a large volume of data (which was not the case).

According to our framework, we got 52 kg of estimated strawberries to be ordered, when only 50 kg were sold. The difference of 2 kg was due to the change of intentions of User-02 and User-07, each of them buying 1 kg less than what they had posted on Facebook.

7.5 Evaluative comparison

None of the models, analyzed in Sect. 4, can be applied to this case study. On one hand, they do not allow the analysis of the three entities of web discourse (user, post and concept) as a whole, and on the other hand, they are not comprehensive enough to address the defined decision problem, nor the data extraction and processing techniques needed for its resolution. This case study not only needs to address the concepts contained in messages, but also the user's network, and therefore, the analyzed models do not meet the necessary conditions to answer questions such as who ordered what and to associate customers with the quantities of ordered strawberries.

The proposed model is better because it can be adjusted to specific contexts, meaning that it does not bear a specific and restricted use of a specific OSN, domain or any decision situation. In addition, it provides a systematic method that allows creating and analyzing various networks, namely of users and semantics. The analysis of semantic data has two purposes: the creation of input data, in order to feed a cleaning database and the answering of questions and decision problems of an organizational context.

Table 8 Quantities of strawberries requested by customer

User	(a) Framework results					(b) Manual results					Δ
	1 kg	2 kg	3 kg	5 kg	Total	1 kg	2 kg	3 kg	5 kg	Total	(a)–(b)
User-01	2	4	0	0	6	2	4	0	0	6	0
User-02	1	2	0	0	3	0	2	0	0	2	1
User-03	0	0	3	0	3	0	0	3	0	3	0
User-04	1	0	0	0	1	1	0	0	0	1	0
User-05	1	2	3	0	6	1	2	3	0	6	0
User-06	2	4	0	0	6	2	4	0	0	6	0
User-07	1	6	0	0	7	0	6	0	0	6	1
User-08	2	2	0	0	4	2	2	0	0	4	0
User-09	1	0	0	0	1	1	0	0	0	1	0
User-10	1	4	0	0	5	1	4	0	0	5	0
User-11	0	2	3	0	5	0	2	3	0	5	0
User-12	0	0	0	5	5	0	0	0	5	5	0
Total	12	26	9	5	52	10	26	9	5	50	2

A first advantage of the proposed model is the fact that it structures the data of users, posts and concepts (the three entities of web discourse), in a linked way, thus allowing the creation of networks that enables a global view, not only of the social interactions among users, but also of the produced dialogs. In this context, some of the models only consider a single entity of analysis, namely the user (Gjoka et al. 2011; Banica et al. 2015; Madan and Chopra 2015; Ghafoor and Niazi 2016). Vosecky et al. (2014), Fernando et al. (2015), Lai and To (2015) only consider the modeling of the semantic content, by using hashtags and web document links to identify keywords. The models that consider all the three entities of web discourse (Oussalah et al. 2013; Vosecky et al. 2014; Walha et al. 2017; Appel et al. 2018) restrict their usage to a specific OSN.

The second advantage is that the proposed model has a three-step structure for extracting, storing, processing and analyzing OSN data. Such steps, which are fundamental for analyzing OSN data, are not fully encompassed in any of the models of Alkhyeli and Mansour (2015), Madan and Chopra (2015), Ghafoor and Niazi (2016). In addition, only the models proposed by Banica et al. (2015) and Vicario et al. (2017) include a direct data extraction from an OSN, while the remaining works merely rely on the use of questionnaires to obtain data.

The third advantage arises from implementing the proposed model by using SNA techniques to visualize the network and to calculate its associated metrics, as well as by making use of NLP (such as data mining and text mining and data cleaning) and common software to support all processes and inherent independence regarding the interpretation of semantic data. Only Zielinski et al. (2013), Banica et al. (2015), Ghafoor and Niazi (2016), Appel et al. (2018) and Ruas et al. (2019) use SNA for user's characterization and NLP. The models by Oussalah et al. (2013), Zielinski et al. (2013), Banica et al. (2015), Lai and To (2015), Vicario et al. (2017), Walha et al. (2017), Alhalabi et al. (2021) and Adikari et al. (2021) depend on external data sources (thesaurus, ontologies, etc.) and/or external software (Leximancer, WordSmith, Stanford parser, etc.) to interpret semantic data. Moreover, some of the models (Oussalah et al. 2013; Zielinski et al. 2013; Vosecky et al. 2014; Fernando et al. 2015; Vicario et al. 2017; Alhalabi et al. 2021; Arafeh et al. 2021) also require complex programming knowledge (namely Twitter or Facebook APIs). Conversely, the proposed model does not depend on external data sources or specific software, using commonly available software (Excel and Gephi) to interpret semantic data.

The fourth advantage emerges from a semantic processing that can use predefined patterns from other systems, when available, but it can also thrive on the concepts directly extracted from the organizational context where it is being applied. This ensures greater reliability of the data, greater control over it and allows to control the vocabulary characteristics of each organizational context. The criteria and restrictions for eliminating and/or correcting data are, thus, defined according to each organizational context. This semantic processing defines more precise standards (without the complexity involved in Semantic Web technologies, LDA, clustering methods, machine learning techniques or ontologies construction), as well as a more accurate analysis of interactions between users and their relevance for each analysis context.

As a last advantage, instead of relying on complex databases such as MySQL, NoSQL, Cassandra, etc., as in Oussalah et al. (2013), Zielinski et al. (2013), Banica et al. (2015), Caroleo et al. (2015), the proposed framework integrates a graph database to organize data in a graph format, making it possible to create and visualize possible sub-networks, according to the purpose of the analysis.

8 Conclusions

In this paper, we presented a framework to extract concepts and summaries of the produced discursive exchanges within an OSN (a Facebook group), to be useful in decision-making. Data were collected, processed and analyzed using SNA. We explored the OSN, combining two different perspectives: the social interactions between users and the semantic analysis of their discourse. To test and illustrate how the framework could be used for decision support, a business idea was chosen, which was implemented with a group of potential customers.

In the case study *sale of strawberries*, presented in this paper, we concluded that it was not only possible to recommend the needed quantities to be transported and delivered to customers, thus optimizing product distribution, but also to identify the customers who requested the most and the leaders of the customers' group. It should, however, be stressed that the framework does not have an implicit purpose of obtaining a quantity. The quantity, in this case, is just an important concept that needed to be made explicit. Therefore, we stand that the presented framework is useful for decision-making in a wide range of applications. In different contexts, different decision goals and concepts can be captured (e.g., addressing the influential buyers; determining causes for product dissatisfaction or complaints; assessing relationships for cross-selling; improving post-sales service; predicting the number of people attending a specific event; etc., are just some of its applications, whether individually or in a combined fashion).

Nonetheless, this work presents context-specific limitations, regarding used linguistics and concepts, that might be different in other settings. Additionally, users from other countries use different languages that might have a different structure. For instance, in English, adjectives precede nouns, while in Latin-based languages, nouns usually precede adjectives. Finally, it is difficult to associate a small *portion of text*, from a post, to a request change, because any post can influence the requested amount. To solve this, we need to improve the semantic data processing algorithms.

In future work, it would be interesting to see comparative studies in other context-specific scenarios and languages, to highlight differences and similarities between semantic extraction procedures. We intend to explore semantic data processing algorithms even further, both in context-specific and context-generic social web networks. Another issue that could be incorporated in the present framework is a trust metric for customers. To do so, literature presents some proposals (although not addressed at this time) of relationship inference algorithms for assigning a trust score to peers. This confidence between nodes (individuals) of a trusted network should make it possible to calculate a weight to be granted to the recommendation of the quantity of strawberries to be transported.

Acknowledgements This work was partially supported by the Portuguese Foundation for Science and Technology under project UIDB/00308/2020 and project UIDB/05037/2020. The authors are also grateful to the anonymous reviewers for their useful comments and suggestions.

References

Abulaish M, Kamal A, Zaki MJ (2020) A survey of figurative language and its computational detection in online social networks. ACM Trans Web 14(1):Article 3

Adib SA, Mahanti A, Naha R (2021) Characterisation and comparative analysis of thematic video portals. Technol Soc 67:121–136

Adikari A, Gamage G, de Silva D, Mills N, Wong S-MJ, Alahakoon D (2021) A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web. Futur Gener Comput Syst 116(2021):302–315

Aggarwal CC (2011) Social network data analytics. Springer, New York

Akar E, Dalgic T (2018) Understanding online consumers' purchase intentions: a contribution from social network theory. Behav Inf Technol 37(5):473–487

Aladwani AM (2014) The 6As model of social content management. Int J Inf Manage 34(2):133–138

Alhajj R, Rokne J (2018) Encyclopedia of social network analysis and mining. Springer, New York

Alhalabi W, Jussila J, Jambi K, Visvizi A, Qureshi H, Lytras M, Malibari A, Adham RS (2021) Social mining for terroristic behavior detection through Arabic tweets characterization. Futur Gener Comput Syst 116(2021):132–144

Alkhyeli M, Mansour A (2015) Using social media for supporting decision-making in managing public relations: the case of Abu Dhabi Police. In: ECSM 2015 2nd European conference on social media. academic conferences and Publishing International, Porto, pp 479–487

Antunes F, Costa JP (2011) Decision support social network conference in information systems and technologies (CISTI). Chaves, Portugal June 15–18

Antunes F, Freire M, Costa JP (2014) Semantic web tools and decision-making (152). In: Zaraté P, Kersten GE, Hernández JE (eds) Group decision and negotiation: a process-oriented view, Lecture Notes in Business Information Processing (LNBIP). Springer, New York, pp 270–277

Appel AP, Santana VFd, Moyano LG, Ito M, Pinhanez CS (2018) A social network analysis framework for modeling health insurance claims data, computer science. Social and Information Networks. arXiv Cornell University. https://arxiv.org/abs/1802.07116

Arafeh M, Ceravolo P, Mourad A, Damiani E, Emanuele B (2021) Ontology based recommender system using social network data. Futur Gener Comput Syst 115(2021):769–779

Arif T (2015) The mathematics of social network analysis: metrics for academic social networks. Int J Comput Appl Technol Res 4(12):889–893

Banerjee S, Jenamani M, Pratihar DK (2017) Properties of a projected network of a bipartite network. In: International conference on communication and signal processing (ICCSP), Chennai, India April 6–8

Banerjee S, Ramanathan K, Gupta A (2007) Clustering short texts using wikipedia. In: Proceedings of the 30th annual international

ACM SIGIR conference on research and development in information retrieval. ACM, USA, pp 787–788
Banica L, Brinzea VM, Radulescu M (2015) Analyzing social networks from the perspective of marketing decisions. Sci Bull Econ Sci 14:37–50
Bapna R, Ramaprasad J, Umyarov A (2018) Monetizing freemium communities: does paying for premium increase social engagement? MIS Q Manag Inf Syst 42(3):719–736
Batagelj V, Doreian P, Ferligoj A, Kejzar A (2014) Understanding large temporal networks and spatial networks: exploration, pattern searching, visualization and network evolution. Wiley, New York
Biswas S, Bordoloi M, Shreya J (2018) A graph based keyword extraction model using collective node weight. Expert Syst Appl 97(2018):51–59
Bouet M, Gançarski P, Aufaure MA, Boussaid O (2009) Pattern mining and clustering on image databases. In: Erickson J (ed) Database technologies: concepts, methodologies, tools, and applications. IGI Global, USA, pp 60–85
Caroleo B, Tosatto A, Osella M (2015) Making sense of governmental activities over social media: a data-driven approach. In: Delibasié B, Hernández JE, Papathanasiou J, Dargam F, Zaraté P, Ribeiro R, Liu S, Linden I (eds) Decision support systems V—big data analytics for decision making. Springer, Serbia, pp 34–45
Chan M, Gong M, Naha R, Mahanti A (2020) Piracy on the internet: publisher-side analysis on file hosting services. In: International symposium on networks, computers and communications (ISNCC), pp 1–7
Chatterjee A, Trumbo BE (2018) Univariate descriptive statistics. In: Alhajj R, Rokne J (eds) Encyclopedia of social network analysis and mining. Springer, New York, pp 3252–3272
Chester S, Kapron BM, Srivastava G, Srinivasan V, Thomo A (2018) Anonymization and de-anonymization of social network data. In: Alhajj R, Rokne J (eds) Encyclopedia of social network analysis and mining, 2nd edn. Springer, New York, pp 78–86
Chua CEH, Storey VC, Li X, Kaul M (2019) Developing insights from social media using semantic lexical chains to mine short text structures. Decis Support Syst 127:1–10
Davis A (2019) Data Wrangling with JavaScript. Manning Publications Company, New York
Duari S, Bhatnagar V (2020) Complex network based supervised keyword extractor. Expert Syst Appl 140:1–14
Erétéo G, Limpens F, Gandon F, Corby O, Buffa M, Leitzelman M, Sander P (2011) Semantic social network analysis: a concrete case. In: Daniel BK (ed) handbook of research on methods and techniques for studying virtual communities: paradigms and phenomena. IGI Global, USA, pp 122–138
Fernando G, MdJohar M, MdJohar M (2015) Framework for social network data mining. Int J Comput Appl Technol Res 116(18):7–10
Freire M, Antunes F, Costa JP (2015) Exploring social network analysis techniques on decision support. In: ECSM 2015 2nd European conference on social media. Porto, pp 165–173
Freire M, Antunes F, Costa JP (2017) A semantics extraction framework for decision support in context-specific social web networks. In: Linden I, Liu S, Colot C (eds) Decision support systems VII data, information and knowledge visualization in decision support systems. Springer, Switzerland, pp 133–147
Fu X, Luo J-D, Boos M (2017) Social network analysis: interdisciplinary approaches and case studies. Taylor & Francis Group, London
Gaeta R (2018) A model of information diffusion in interconnected online social networks. ACM Trans Web 12(2):Article 13
Ghafoor F, Niazi MA (2016) Using social network analysis of human aspects for online social network software: a design methodology. Complex Adapt Syst Model 4(14):1–19
Ghim G-H, Kim K, Ko Y, Bae S, Choi W (2018) NetMiner. In: Alhajj R, Rokne J (eds) Encyclopedia of social network analysis and mining, 2nd edn. Springer, New York, pp 1450–1474
Gjoka M, Kurant M, Butts CT, Markopoulou A (2011) A walk in Facebook: uniform sampling of users in online social networks. In: IEEE INFOCOM (2010). IEEE journal on selected areas in communications (JSAC), San Diego, pp 1–9
Golbeck J (2015) Introduction to social media investigation: a hands-on approach. Elsevier, New York
Hanneman RA, Riddle M (2005) Introduction to social network methods. 2005. http://faculty.ucr.edu/~hanneman/networks/nettext.pdf. Accessed 12 Apr 2015
Harrison G (2015) Next generation databases NoSQL, NewSQL, and big data. Springer, New York
Herring S (2013) Discourse in Web 2.0: familiar, reconfigured, and emergent. In: Tannen D, Trester A-M (eds) Discourse 2.0: language and new media. Georgetown University Press, Washington, pp 1–25
Himelboim I, McCreery S, Smith M (2013) Birds of a feather tweet together integrating network and content analyses to examine cross-ideology exposure on Twitter. J Comput Mediat Commun 18:154–174
Ignatov DI, Khachay MY, Labunets VG, Loukachevitch N, Nikolenko SI, Panchenko A, Savchenko AV, Vorontsov K (2017) Analysis of images, social networks and texts. Springer, New York
Ikematsu K, Murata T (2013) A fast method for detecting communities from tripartite networks. In: Jatowt A, Lim E-P, Ding Y, Miura A, Tezuka T, Dias G, Tanaka K, Flanagin A, Dai BT (eds) Social informatics. Springer, Switzerland, pp 192–205
Ishfaq U, Khan HU, Iqbal S, Alghobiri M (2021) Finding influential users in microblogs: state-of-the-art methods and open research challenges. Behav Inf Tech 1–44 (**ahead-of-print**)
Isson JP (2018) Unstructured data analytics: how to improve customer acquisition, customer retention, and fraud detection and prevention. Wiley, New York
Kemper C (2015) Beginning Neo4j create relationships and grow your application with Neo4j. Springer, New York
Kleminski R, Kazienko P (2018) Identifying promising research topics in computer science. In: Alhajj R, Hoppe HU, Hecking T, Brodka P, Kazienko P (eds) Network intelligence meets user centered social media networks. Springer, Switzerland, pp 231–241
Kok S, Rogers R (2017) Rethinking migration in the digital age-transglocalization and the Somali diaspora. Glob Netw 17(1):23–46. https://doi.org/10.1111/glob.12127
Lai LSL, To WM (2015) Content analysis of social media: a grounded theory approach. J Electron Commer Res 16(2):138–152
Liu W, Xiaojun Q, Min F, Bite Q (2010) A short text modeling method combining semantic and statistical information. Inf Sci 180(20):4031–4041
Liu X, Min Q, Wu D, Liu Z (2020) How does social network diversity affect users' lurking intention toward social network services? A role perspective. Inf. Manag. 57(7):1–16
Lukanin A (2015) Normalization of non-standard words with finite state transducers for russian speech synthesis. In: Khachay MY, Konstantinova N, Panchenko A, Ignatov DI, Labunets VG (eds) Analysis of images social networks and texts. Springer, Switzerland, pp 39–48
Ma H, Che D (2016) An integrative social network and review content based recommender system. J Ind Intell Inf 4(1):69–75
Madan M, Chopra M (2015) Using mining predict relationships on the social media network: Facebook (FB). Int J Adv Res Artif Intell 4(4):60–63
Mahanti A, Carlsson N, Williamson C (2012) Content sharing dynamics in the global file hosting landscape. In: Conference in proceedings of the 2012 IEEE 20th international symposium on modeling,

analysis and simulation of computer and telecommunication systems. https://doi.org/10.1109/MASCOTS.2012.34
Marmo R (2011) Web mining and social network analysis. In: Zhang H, Segall R, Cao M (eds) Visual analytics and interactive technologies: data, text and web mining applications. Information Science Reference, Hershey, pp 202–211
Monaghan S, Lavelle J, Gunnigle P (2017) Mapping networks: exploring the utility of social network analysis in management research and practice. Elsevier J Business Res 76(C):136–144
Moser C, Groenewegen P, Huysman M (2013) Extending social network analysis with discourse analysis: combining relational with interpretive data. In: Özyer T, Rokne J, Wagner G, Reuser A (eds) The influence of technology on social network analysis and mining. Springer, New York, pp 547–561
Navigli R, Lapata M (2010) An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Trans Pattern Anal Mach Intell 32(4):678–692
Opsahl T (2013) Triadic closure in two-mode networks: redefining the global and local clustering coefficients. Elsevier Soc Netw 35:159–167
Oussalah M, Bhat F, Challis K, Schnier T (2013) A software architecture for Twitter collection, search and geolocation services. Knowl Based Syst 37:105–120
Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Technical report, Stanford University, Stanford, CA
Pang B, Lee L (2008). Opinion mining and sentiment analysis. foundations and trends in information Retrieval, 2008. http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf
Power DJ, Phillips-Wren G (2012) Impact of social media and Web 2.0 on decision-making. J Decis Syst 20(3):249–261
Provost F, Fawcett T (2013) Data science for business: what you need to know about data mining and data-analytic thinking. O'Reilly Media, NewtonA
Réda S, Baba-Hamed L, Djatou A (2017) Twitter social networking for recommendation. In: Linden I, Mareschal B, Shaofeng L, Papathanasiou J, Colot C (eds) ICDSST 2017—3rd international conference on decision support system technology: data, information and knowledge visualization in decision making. ICDSST, Namur, Belgium, pp 161–168
Robinson I, Webber J, Eifrem E (2015) Graph databases. O'Reilly Media Inc, Gravenstein Highway North, Sebastopol
Rosa KD, Shah R, Lin B, Gershman A, Frederking R (2011) Topical clustering of tweets. In: Proceedings of SWSM'10. Beijing, China
Ruas PHB, Machado AD, Silva MC, Meireles MRG, Cardoso AMP, Zárate LE, Nobre CN (2019) Identification and characterisation of Facebook user profiles considering interaction aspects. Behav Inf Technol 38(8):858–872
Saint-Charles J, Mongeau P (2018) Social influence and discourse similarity networks in workgroups. Elsevier Soc Netw 52:228–237
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web. ACM, USA, pp 851–860
Salmons J (2017) Using social media in data collection: designing studies with the qualitative e-research framework. In: Sloan L, Quan-Haase A (eds) The SAGE handbook of social media research methods. SAGE Publications Ltd, London, pp 177–196
Samanthula BK, Jiang W (2014) A Randomized approach for structural and message based private friend recommendation in online social networks. In: Can F, Ozyer T, Polat F (eds) State of the art applications of social network analysis. Springer, Switzerland, pp 1–34
Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T, Gonzalez G (2015) Utilizing social media data for pharmacovigilance: A review. J Biomed Inf 54:202–212
Sathick J, Venkat J (2015) A generic framework for extraction of knowledge from social web sources (social networking websites) for an online recommendation system. Int Rev Res Open Distrib Learn 16(2):247–271
Savic M, Ivanovic M, Jain LC (2019) Complex networks in software, knowledge, and social systems. Springer, Switzerland
Simon HA (1977) The new science of management decision. Prentice Hall, Englewood Cliffs
Tabassum S, Gama J, Azevedo P, Teixeira L., Martins C, Martins A (2021) Dynamic topic modeling using social network analytics. Cham
Tollinen A, Jarvinen J, Karjaluoto H (2012) Opportunities and challenges of social media monitoring in the business to business sector. In: The 4th international business and social science research conference. Dubai, UAE, pp 1–14
Tripathy BK, Baktha K (2018) Security, privacy, and anonymization in social networks: emerging research and opportunities: emerging research and opportunities. IGI Global, USA
Tripathy BK, Thakur S, Chowdhury R (2017) A Classification Model to Analyze the Spread and Emerging Trends of the Zika Virus in Twitter. In: Behera HS, Mohapatra DP (eds) Computational intelligence in data mining. Springer, Singapore, pp 643–650
Troisi O, Grimaldi M, Loia F, Maione G (2018) Big data and sentiment analysis to highlight decision behaviours: a case study for student population. Behav Inf Technol 37(10–11):1111–1128
Velde BVD, Meijer A, Homburg V (2015) Police message diffusion on Twitter: analysing the reach of social media communications. Behav Inf Technol 34(1):4–16
Vicario M, Zollo F, Caldarelli G, Scala A, Quattrociocchi W (2017) Mapping social dynamics on Facebook: the Brexit debate. Soc Netw 50:6–16
Vosecky J, Jiang D, Leung KW-T, Xing K, Ng W (2014) Integrating social and auxiliary semantics for multifaceted topic modeling in Twitter. ACM Trans Internet Technol (TOIT) Special Issue Found Soc Comput 14(4):1–24
Wachsmuth H (2015) Text analysis pipelines: towards ad-hoc large-scale text mining. Springer, Switzerland
Walha A, Ghozzi F, Gargouri F (2017) ETL4Social-data: modeling approach for topic hierarchy. In: 9th international joint conference on knowledge discovery, knowledge engineering and knowledge management (KEOD 2017), Madeira, Portugal November 1–3
Wang L, Ren X, Wan H, Yan J (2020) Managerial responses to online reviews under budget constraints: whom to target and how. Inf Manag 57(8):1–13
Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, New York
Yu R, Christophersen C, Song Y-D, Mahanti A (2019) Comparative analysis of adult video streaming services: characteristics and workload. In: 2019 network traffic measurement and analysis conference (TMA), pp 49–56
Zielinski A, Middleton SE, Tokarchuk LN, Wang X (2013) Social media text mining and network analysis for decision support in natural crisis management. In: International conference on information systems for crisis response and management (ISCRAM), Baden-Baden, Germany May 12–15