Revista Gestão & Tecnologia
e-ISSN: 2177-6652
revistagt@fpl.edu.br
http://revistagt.fpl.edu.br/

OPINION MINING FRAMEWORK APPLIED TO A SOCIAL NETWORKS DATA FOR SMALL AND MEDIUM ENTERPRISES

FRAMEWORK DE MINERAÇÃO DE OPINIÕES APLICADO A DADOS DE REDES SOCIAIS PARA PEQUENAS E MÉDIAS EMPRESAS

FRAMEWORK PARA LA MINERÍA DE OPINIÓN APLICADO A DATOS DE LAS REDES SOCIALES PARA LAS PEQUEÑAS Y MEDIANAS EMPRESAS

Huoston Rodrigues Batista
Doutorando em Informática e Gestão do Conhecimento pela Universidade Nove de Julho
huoston@uni9.pro.br
https://orcid.org/0000-0002-8801-5770

Marco Antonio Gaspar
Doutor em Administração pela USP, docente permanente e pesquisador do Programa de Pós-graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho
marcos.antonio@uni9.pro.br
https://orcid.org/0000-0002-2422-2455

Renato José Sassi
Doutor em Engenharia Elétrica pela USP, docente permanente e pesquisador do Programa de Pós-graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho
sassi@uni9.pro.br

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 27.03.2019
Aprovado em 02.08.2020

Abstract

Title: Opinion mining framework applied to a social networks data for small and medium enterprises.

Objective: To present a framework for the mining of opinions that can be applied in the discovery of knowledge of the customers about to their experiences, based on unstructured data extracted from social networks, and that is applicable to the reality of small and medium enterprises.

Methodology: This experimental research accessed data from the opinions of customers of four restaurants published in the social network TripAdvisor Brazil. The framework was based on the proposals formulated by Aranha (2007) and Feldman and Sanger (2007), techniques for Sentiment Analysis by Liu (2012) and Pang and Lee (2008) and Topic Modeling by Blei *et al.* (2012).

Originality: The relevance consists in proposing a solution that is both accessible to SMEs and capable of processing opinions in Portuguese, something not very common in literature. Almost all similar applications in literature are dedicated to the English language.

Main results: We highlight the generation of summaries and graphic visualizations that contribute to evidence knowledge about the relations between several expressions and terms that were not obvious. These allowed finding latent relationships between terms cited by different customers.

Theoretical contributions: The methodological solution uses efficient and state-of-the-art techniques and methods to extract, process, and analyze customer opinions on the Internet quickly, efficiently, and economically.

Social contributions: the framework developed presents an efficient, fast and economical way to mine data, presenting the results of the discovery of customer knowledge through the use of Sentiment Analysis and Topic Modeling techniques.

Keywords: Data mining; Text mining; Opinion mining; Social networks; Customer knowledge.

Resumo

Título: *Framework* de mineração de opiniões aplicada a dados de redes sociais para pequenas e médias empresas.

Objetivo: Apresentar um *framework* para a mineração de opiniões que possa ser aplicado na descoberta de conhecimento de clientes sobre suas experiências, a partir de dados não estruturados extraídos das redes sociais, e que seja aplicável à realidade de pequenas e médias empresas.

Metodologia: Esta pesquisa experimental acessou dados das opiniões de clientes de quatro restaurantes publicadas na rede social TripAdvisor Brasil. O *framework* desenvolvido baseou-se nas propostas de Aranha (2007) e Feldman e Sanger (2007), técnicas de Análise de Sentimento de Liu (2012) e Pang and Lee (2008) e Modelagem de Tópicos de Blei et al. (2012).

Originalidade: A relevância consiste em propor uma solução acessível às PMEs e capaz de processar opiniões em português, algo pouco comum na literatura. Quase todas as aplicações similares disponíveis na literatura voltam-se à língua inglesa.

Principais resultados: Destaca-se uma solução para a geração de resumos e visualizações gráficas que contribuem para evidenciar o conhecimento sobre as relações entre diversas expressões e termos que não são óbvias. Isso permitiu encontrar relações latentes entre termos citados por diferentes clientes.

Contribuições teóricas: A solução metodológica utiliza técnicas e métodos eficientes e de última geração para extrair, processar e analisar opiniões dos clientes na Internet de forma rápida, eficiente e econômica.

Contribuições sociais: o *framework* desenvolvido apresenta uma forma eficiente, rápida e econômica de minerar dados, apresentando os resultados da descoberta do conhecimento do cliente por meio da aplicação de técnicas de Análise de Sentimento e Modelagem de Tópicos.

Palavras-chave: Mineração de dados; Mineração de texto; Mineração de opinião; Redes sociais; Conhecimento do cliente.

Resumén

Título: *Framework* para la minería de opinión aplicado a datos de las redes sociales para las pequeñas y medianas empresas

Objetivo: Presentar un *framework* para la minería de opiniones que pueda aplicarse en el descubrimiento del conocimiento de los clientes a sus experiencias, basado en datos no estructurados extraídos de redes sociales, y que sea aplicable a la realidad de los PYME.

Metodología: Esta investigación experimental accedió a los datos de las opiniones de los clientes de cuatro restaurantes publicadas en TripAdvisor Brasil. El *framework* se basó en las propuestas de Aranha (2007) y Feldman; Sanger (2007), técnicas para el análisis de sentimientos de Liu (2012) y Pang; Lee (2008) y Topic Modeling de Blei *et al.* (2012).

Originalidad: La pertinencia consiste en proponer una solución accesible a las PYME y capaz de tramitar dictámenes en portugués, algo que no es muy común en la literatura. Casi todas las aplicaciones similares en la literatura están dedicadas al idioma inglés.

Principales resultados: Generación de resúmenes y gráficos que contribuyen a evidenciar el conocimiento sobre las relaciones entre varias expresiones y términos que no eran obvios. Esto permitió encontrar relaciones latentes entre los términos citados por diferentes clientes.

Contribuciones teóricas: La solución metodológica utiliza técnicas y métodos eficientes y de última generación para extraer, procesar y analizar las opiniones de los clientes en Internet de manera rápida, eficiente y económica.

Contribuciones sociales: el *framework* presenta una forma eficiente, rápida y económica de extraer datos, presentando los resultados del descubrimiento del conocimiento del cliente mediante el uso de técnicas de Análisis de Sentimientos y Modelado de Temas.

Palabras clave: Minería de datos; Minería de texto; Minería de opinión; Redes sociales; Conocimiento del cliente.

1. Introduction

In the current competitive environment in which companies are inserted, promoting the interaction between the company's resources and knowledge is vital to its success. There is no faster or more interactive means of communication than the internet (Kaplan & Haenlein, 2010; Kietzmann, Hermkens, McCarthy, & Silvestre, 2011). Customers can connect with businesses and other customers through social networks more interactively than ever before, enabling companies to deepen their relationships with customers.

Low cost, personalization and ease of content creation focused on messages through social media present themselves as relevant advantages over traditional communication channels (Hoffman & Fodor, 2010). Such context makes the use of social media not only relevant, but a strategic factor for companies, regardless of their size or segment (Ammirato et al., 2019). However, using social media is not an easy task and may require new ways of thinking by companies (Kaplan & Haenlein, 2010).

These characteristics of the current business environment make social media a particularly useful medium for Small and Medium Enterprises (SMEs), a segment known both for its resource limitations and for its potential, since this type of business often does not have management of its business areas and do not normally invest in administrative or customer relationship tools (Lin, 2014; Esposito & Evangelista, 2016).

In Brazil, SMEs account for 27% of national GDP (Gross Domestic Product), 52% of jobs with a formal contract and 40% of salaries paid, according to SEBRAE - Brazilian Service of Support to Micro and Small Enterprises latest statistics released (Sebrae, 2014). Such participation is relevant and justifies the concern with the performance and potential evolution of the companies included in this particular segment, as well as its main asset: the customer.

Due to the presented context, the knowledge of the customer has aroused the interest of researchers as a discipline and, moreover, as a fundamental strategic source for the success of any company (Rowley, 2002; Campbell, 2003; Rollins & Halinen, 2005).

However, knowledge, as an important resource for contemporary companies, often comes from data and information that need to be processed in order to take on new

configurations that convey meaning, thus contributing for decision making in the enterprise.

With the technological evolution marked by the connectivity and popularization of mobile devices and, above all, by the increasing penetration of the Internet in all aspects of modern society, there is a significant increase in the volume of data. Such increase in volume mentioned comes predominantly from unstructured data, such as images, videos and, mainly, texts published in social networks (L. A. da Silva, Peres, & Boscarioli, 2017).

Therefore, it is necessary to develop techniques and methods that make it possible to extract knowledge of these data, a task that, due to its complexity, motivated the emergence of different methods and techniques for performing data mining from the complex web of social network repositories.

This research aims to present a framework for the mining of opinions that can be applied in the discovery of knowledge of the customers about to their experiences, based on unstructured data extracted from social networks, and that is applicable to the reality of small and medium enterprises.

2. Literature Review

2.1. Customer Knowledge

From 1990's on, knowledge has become the most important and indispensable resource for any company that seeks to survive in an increasingly connected world, where products, services and values circulate in ways never before imagined, thus being more valuable and powerful than any other physical or financial (Stewart, 1997). Considering the importance of the concept, it is necessary to explore some of its main definitions, as well as their applicability in the context of companies. Perhaps because it is a complex concept, knowledge is defined in many ways in the literature.

One of the most cited definitions of knowledge in the literature is the process of transforming data into information and information into knowledge (Davenport & Prusak, 2000). The authors define that this transformation takes place through comparisons, consequences, connections and conversations.

Knowledge differs from information because it is about beliefs, commitment and action (Nonaka & Takeuchi, 1995), although the authors cite that knowledge and information share something in common: both should be imbued with meaning.

With regard to the nature of knowledge, it can be divided into two great dimensions: tacit and explicit (Polanyi, 1966; Nonaka & Takeuchi, 1995; Dalkir & Liebowitz, 2011). Explicit knowledge is knowledge that has somehow been captured, stored and subsequently made available for sharing. Explicit knowledge, being expressed in words, numbers, is communicated and shared in the form of raw data, scientific formulas, codified processes or universal principles (Nonaka & Takeuchi, 1995).

On the other hand, tacit knowledge is a somewhat more complex form of knowledge (Polanyi, 1966). According to Nonaka and Takeuchi (1995), Tacit knowledge is something extremely personal and therefore difficult to formalize, making it difficult to communicate or share with others. Moreover, tacit knowledge is deeply rooted in the actions and experiences of individuals, as well as in the ideas, values, beliefs and emotions that people adopt.

According Behringer, Sassenberg & Scholl (2017), knowledge exchanges via social media are critical for organizational success nowadays. Crammond & Murray (2018) say that the use of social media, especially in small and medium companies, is contributory to management knowledge in the firm. In fact, in the last two decades customer knowledge had its value recognized by scholars as a discipline and, moreover, as a key strategic resource for the success of any company (Rowley, 2002; Rollins & Halinen, 2005). In addition, customer knowledge is pointed out in the literature as a way to support the long-term relationship with customers (Darroch & McNaughton, 2003). According to Campbell (2003) customer knowledge refers to the understanding of the customer's needs, expectations and objectives, and is an essential component for any company that intends to build real relationships with its customers.

The customer can provide unique insight that enables companies to learn and improve their internal operations (Paquette, 2011). However, Meneghello *et al.* (2019) indicate that due to collection and extraction challenges, data in many feeds, embedded comments, reviews and testimonials are quite inaccessible as a generic data source.

The relationship between the company and the customer is described as a dynamic process, where both change over time (Nejatian, Sentosa, Piaralal, & Bohari, 2011), which in turn may require an effort on the part of the company not only to approach this client, but also to try to extract from this his opinions and impressions on his experiences of constant form. On the other hand, better understanding of the client makes it possible to understand their real needs and desires as well, which can significantly improve the relationship between the company and its consumers (García-Murillo & Annabi, 2002). Furthermore, Chierici *et al.* (2019) state that

customers' data gathered from social media produce different effects on knowledge management practices of the company.

However, this is not a trivial task, and requires not only a great willingness on the part of companies but also the use of technologies to extract and understand what is behind these opinions. In this work, we will present a way of extracting, processing and visualizing knowledge from the opinions of clients coming from social networks using, for this, text mining techniques, sentiment analysis and modeling topics.

2.2. Social Media

Web 2.0 is a broad term, coined initially in 2005 by Tim O'Reilly (O'reilly, 2005) and used since then to describe a variety of web-based applications. Web 2.0 is not a new version of the web, but refers to new ways of using the Internet to generate content, explore connections between users, and encourage participation and transparency (O'reilly, 2005).

Some applications and tools commonly referred as Web 2.0 include blogs, podcasts, RSS feeds, social networks, photo and video sharing services, wikis, shareable favorite content environments, and peer-to-peer content distribution services (O'reilly, 2005; Kaplan & Haenlein, 2010).

According to Kaletka and Pelka (2011), Web 2.0 can be considered one of the most influential and important innovations in the field of Information and Communication Technology, responsible for originating platforms worldwide consecrated as innovations in themselves, such as Wikipedia, YoutTube and social media like Facebook and Twitter.

The concept of social media is very broad and sometimes without consensual definition in literature. Correa *et al.* (2010) define social media as a form of specific consumption of digital content. Kaplan and Haenlein (2010) defines social media as a set of Internet applications built on the ideological and technological foundations of Web 2.0, that allows the creation and exchange of user-generated content.

While these definitions cite social media users quite broadly, social media is not confined to personal use. Kietzmann *et al.* (2011) cite the use of social media by companies that seek to increase the company's financial return and improve the brand image through its online presence. The authors defend the idea that such technologies are currently fundamental to the survival of contemporary companies. Nisar, Prabhakar and Strakova (2019) state that

information from social media benefits knowledge management of the company, providing a potent means for organization establishes performance improvements.

Social networking has created opportunities to establish interaction between people and small companies that lead to the sharing of valuable knowledge (Mamorobela & Buckley, 2018). The problem is that this knowledge is almost always available in the form of unstructured and imprecise data by nature. To deal with this, one can apply text mining techniques, which will also deal with Natural Language Processing, fundamental concepts for the development of the opinion mining framework proposed by this research.

2.3. Text Mining

When people communicate, they do it in many ways: they write books, articles, blogs and web pages, interact by sending messages in different ways and, of course, talk to one another. When this happens electronically, this text data becomes a significant resource, which has enormous potential value for a wide range of organizations (Feldman & Sanger, 2007).

Due in particular to the advancement of Web 2.0 technologies, the volume of data and the speed with which they can be queried are higher with each passing day. However, people's ability to process and understand these data remains constant (Hofmann & Chisholm, 2013). To overcome human limitation, the area of Natural Language Processing (NLP) is dedicated to investigate, propose and develop computational systems that have written natural language as object of study (Feldman & Sanger, 2007; Berry & Kogan, 2010). In a simpler way, Natural Language Processing can be defined in a broad sense as any kind of manipulation of natural language into a computer, which may involve a variety of methods and techniques (Clark, Fox, & Lappin, 2010).

The process of extracting value from text data, known as Text Mining, is one of the areas of knowledge that has attracted the attention of market professionals and academics researchers in recent years, especially due to the evolution of the internet and the media. mobile communication (Berry & Kogan, 2010). Several techniques and processes were developed to retrieve relevant information contained in unstructured databases, giving rise to the area known as Text Mining, which, in turn, derives from a wider area known as Data Mining (Feldman & Sanger, 2007).

Text Mining is essentially a multidisciplinary activity and, in addition to using certain techniques of Data Mining, it is also supported in other areas, such as Computational

Intelligence, Information Retrieval, Cognitive Science and, above all, Natural Language Processing (Chowdhury, 2003).

There are in the literature a series of approaches in relation to the processes of Text Mining. Aranha (2007) presents a complete model for acquiring knowledge from texts that consist of four main steps: (1) data extraction (also referred to as document collection); (2) pre-processing; (3) pattern extraction; and (4) analysis and evaluation of results (Aranha, 2007).

Similarly to the Aranha (2007) proposal, Feldman and Sanger (2007), state that, on a functional level, text mining systems follow the general model provided by some classical data mining applications and therefore are segregated into four main areas: (1) Preprocessing tasks; (2) Mining operations; (3) Presentation layer; and (4) Refinement tasks.

For the context of this research, the framework of opinion mining proposed is based on the works of Feldman and Sanger (2007) and Aranha (2007). This choice is due to two reasons: firstly, they are universal in character, that is, they are very similar to several other proposals, which vary in relation to the scope or objective and may contain more specific steps or processes. The second reason is that neither of them would be able to fully contemplate the objectives proposed by this research.

Therefore, this work suggests the adoption of parts of the two models and the addition of specific processes that aim to meet the objectives proposed in this research.

2.4. Opinion Mining

Opinion Mining, also known as The Sentiment Analysis, is an area of Computer Science that is dedicated to the problem of identifying emotions and opinions in textual contents (Pang & Lee, 2008; Liu, 2012). Manivannan and Selvi (2019) state that opinion mining is a process of extorting the opinions of customer about a service, product and policy. The Opinion Mining is often approached in the literature as an activity of classification of sentiments, or, classification of polarity of sentiments, that is, positive, negative or neutral (Liu, 2012).

Opinion Mining is directly related to a specific domain, although there are approaches that address the issue of domain independently. Walaa Medhat, Ahmed Hassan and Hoda Korashy (2014) conducted extensive literature review on the different applications and techniques applied to Opinion Mining, as shown in Figure 1, with the path followed by this research highlighted in blue.

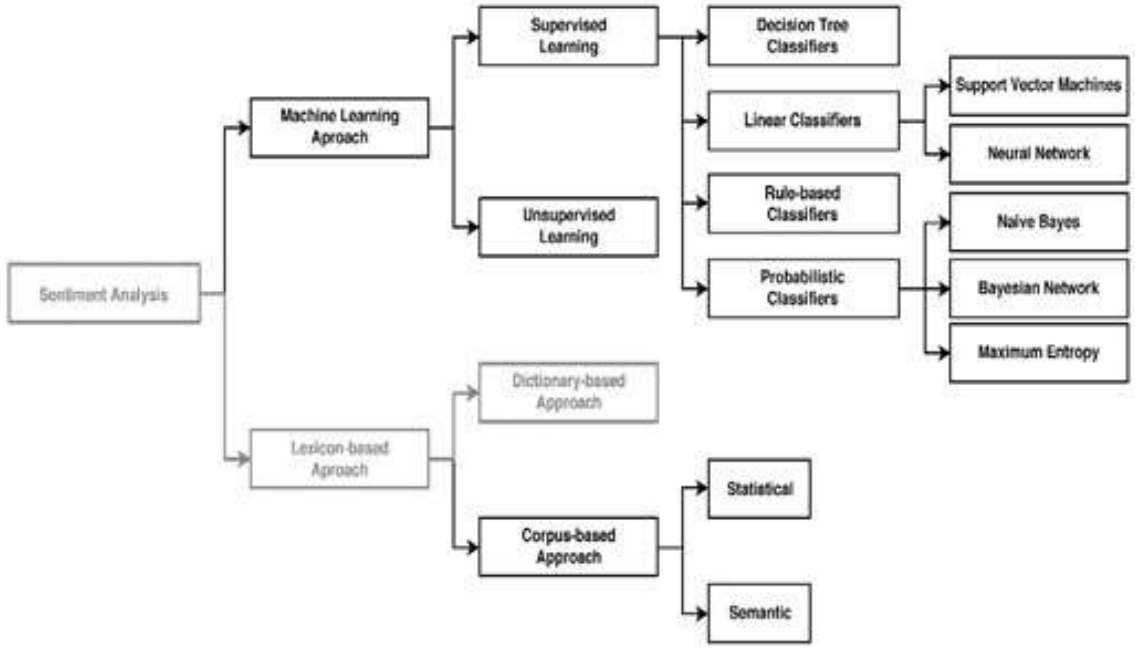

Figure 1 - Approaches applied in the Opinion Mining

Source: adapted from Medhat *et al.* (2014).

For the purposes of this work, Opinion Mining is applied in order to understand the inclination of customer reviews about their experiences in restaurants, area addressed in this research. The approach chosen was based on lexicon and this choice is due to a number of factors, particularly the limitations related to the amount of documents to be analyzed.

The choice is based on the assumption that because the firms involved in this research are Small or Medium, the amount of feedback may not be large enough for the application of techniques that use machine learning approaches, since such techniques require a relatively large amount of data for training and testing activities. Therefore, because of this limitation, statistical or dictionary-based methods tend to be more recommended (Pang & Lee, 2008; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011).

In addition to understanding the general sentiment of customers' opinions regarding their experiences with the restaurants covered in this research, it is necessary to know what customers are most talking about when they praise or criticize some aspect. To achieve this end, we opted for the technique of topic modeling as a way of extracting subjects about which most people talk, discussed below as a way of extracting knowledge from the opinions of customers extracted from the social networks of the companies that constitute the object this research.

2.5. Topic Modeling

The current scale of content generation in the form of texts and their wide availability of access have created demands for organize and classify this type of data that could not be supplied by human annotation. Due to this limit, a possible solution to handle such a volume of data is based on probabilistic topic modeling techniques, whose main objective is the discovery of topics and the annotation of large collections of documents by thematic classification (Blei, 2012).

Such techniques employ statistical methods to the words of the original texts in order to discover the themes present in them, the relation of these themes to each other and how they evolve over time. Topic modeling algorithms do not require any labeling or prior classification of documents, and the topics emerge from the analysis of the original texts as they were produced (Liu, 2012; Blei, 2012).

The Latent Dirichlet Allocation (LDA) is a generative probabilistic model and mixed association model that was introduced by Blei *et al.* (2012), which discovers latent topics in text bodies. Since then, it has become a widely used model in Natural Language Processing. Figure 2 presents the intuitions behind Latent Dirichlet Allocation.

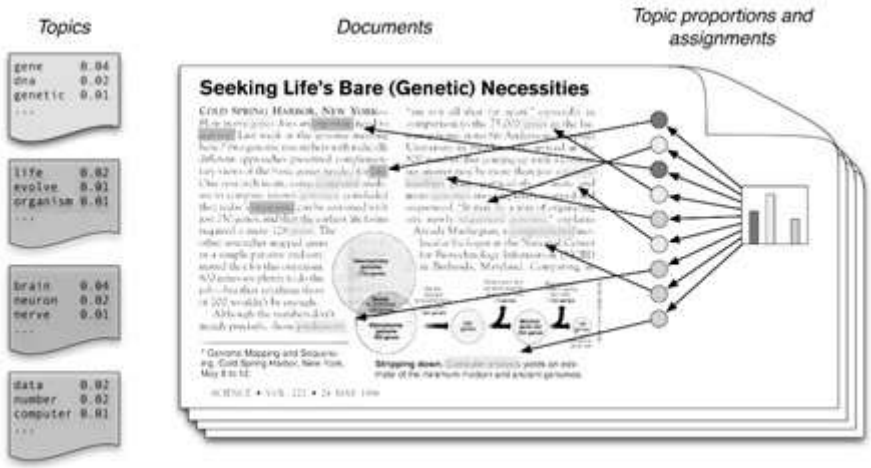

Figure 2 - Assignment of topics to a document in LDA.

Souce: adapted from Blei (2012).

The LDA model discovers the latent topic structure by reversing the formalized generative process. From the information observed - that is, the co-occurrence patterns in the

word distribution of the *corpus* document set - the model infers the outline and its distribution from the generative model (Blei, 2012).

The result of the application of this technique allows the exploration of the latent topic structure present in a corpus. In this work, LDA is applied to extract knowledge from customer reviews, returning a number of topics and the most likely terms in each of the topics.

3. Method and Experimental Design

In this topic we analyze and discuss the stages of development of the proposed framework for discovery of customer knowledge related to their experiences in restaurants. To achieve this purpose, the experiment used data extracted from social networks, applicable to the reality of small and medium enterprises.

3.1. Framework for Opinion Mining

Because it is a developing area, many researchers suggest different methodologies and approaches for text mining application. Thus, there is no consensus as to which steps should be implemented or not, with some arguing that the decision on which steps to take should be evaluated in the context of each specific application (Aranha, 2007; Feldman & Sanger, 2007; L. A. da Silva et al., 2017).

Considering the objective of this research, it is presented in Figure 3 the proposal of a framework for opinion mining that seeks to discover knowledge from the opinions of customers of restaurants published in the social network TripAdvisor Brazil. This social network chosen for this work contains opinions of users about their experiences in restaurants and other types of business.

The restaurants chosen for the development of this research belong to the segment of fast food (harburguers, more specifically), and, as mentioned previously, that fit the definition of Small and Medium Enterprise. The choice of the type of restaurant considered in this research is justified by the predominantly young target public and, therefore, more prone to the use of social networks (Casey, 2017).

The Opinion Mining Framework presented is based on the proposals formulated by Aranha (2007) and Feldman and Sanger (2007), extending them with the application of

additional techniques for Sentiment Analysis (Liu, 2012; Pang & Lee, 2008) and Topic Modeling (Blei *et al.*, 2012).

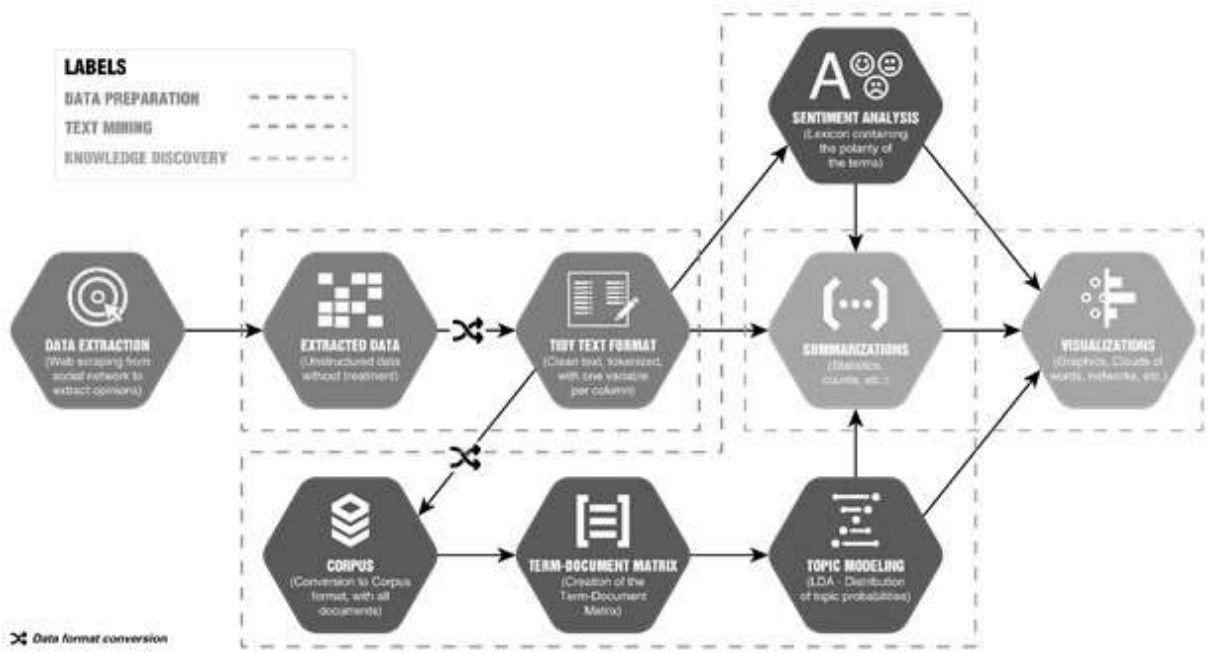

Figure 3 - Opinion mining framework for customer knowledge discover.
Source: authors.

The use of Sentiment Analysis and Topic Modeling is justified in this research in function of the general objective outlined, in this case, the extraction of knowledge from the customers opinions through the application of Opinion Mining techniques. The opinion mining framework adopted for this research relies on a data mining approach known as *Tidy Data* (Wickham, 2014), which is nothing more than the application of a series of principles. This seeks to ensure the organization of values in a data set, and which consists of three rules: each variable forms a column; each observation forms a line and each type of observation unit forms a table.

Silge and Robinson (2017) define the tidy text format as a table with one token per line. The token is a meaningful unit of text (usually a word) that you are interested in analyzing. Thus, tokenization is the process of dividing a text into tokens (Feldman & Sanger, 2007; L. A. da Silva et al., 2017). This structure by a token line contrasts with the ways in which the text is normally stored in the text analysis: as strings or a document-term matrix (Feldman & Sanger,

2007; L. A. da Silva et al., 2017). To perform tidy text mining, the token that is stored in each line is usually a single word, but can also be an n-gram, a phrase, or a paragraph.

In order to implement the presented framework presented in Figure 3, we used tools that perform: (1) the extraction of the opinions of users of restaurants published in the social network TripAdvisor Brasil, (2) the preprocessing of texts in order to format and organize in a way suitable for text mining and (3) text mining and display of graphical visualizations. For the development of the experiments of this research we chose the use of the R language for its flexibility and ease of use, besides a great offer of libraries with algorithms for implementation of the several steps required by the proposed framework.

Regarding the packages that are part of the tools applied in the experiments of this research, we highlight the application of the following: the *reshape2* and *stringr* packages used for data conversion, *tibble* package, consisting of a modern data.frame format, the magrittr package that allows concatenating commands and executing them in sequence, the tidytext package that has been massively applied in text mining operations, the topicmodels package, which includes algorithms for topic modeling and the *ggplot2*, *wordcloud*, *igraph* and *ggraph* packages, used to generate the various graphical visualizations of data mining results. In addition, the *rvest* package was used in the data extraction phase of the selected social network.

3.2. Data analysis

In this research we opted for the TripAdvisor social network, specialized in receiving user opinions about their experiences with enterprises such as hotels, restaurants and events. The focus of this research turned specifically to small restaurants, having addressed three establishments specializing in the same type of product: burgers. Due to space limitations, this article presents the analysis of data from only one of the four companies covered in this research (which possessed greater amount of opinions) hereinafter referred to as COMPANY, described below. COMPANY is a restaurant specializing in traditional American burgers.

Analyses begin shortly after the data preprocessing step, which consists of token creation and removal of stop words and special characters and numbers.

Table 1 shows the result of COMPANY data mass after each preprocessing stage.

Table 1

Summarization of COMPANY data after pre-processing procedures.

Procedure	Number of items after procedure
Input	1.275 documents
Tokenization	58.668 words
Removal of stop words	34.083 words
Removing numbers and special characters	33.716 words

Source: authors.

As a result of the pre-processing stages, a set of 33,716 words without stop words was distributed in the 1,275 documents of COMPANY. Analyzing the repetitions of words (the ten most frequent), it is noticed that by the absolute repetition of the terms in the corpus, customers talk a lot about food (1058 repetitions), location (739 repetitions), and service (442 repetitions). Figure 4 presents the word cloud that allows to evaluate the presence of certain terms regarding how these terms are repeated in the corpus of COMPANY.

Figure 4 - Cloud of words with the most frequent terms of the COMPANY[1].

Source: authors.

The word cloud shown in Figure 4 indicates that the most frequent terms in this corpus refer, mostly, to the service (atendimento), the place (ambiente) and the food (lanches), as they apparently have the same size ratio. The more peripheral and orbiting terms center around a number of interesting aspects, including items such as 'menu' and 'decoration' (decoração), but also reveal occurrences that deserve greater attention, such as the term 'wait' (espera), for example. Given its size, it is probably repeated in several documents. The next step is to perform the sentiment analysis in order to evaluate the inclinations of the customers' opinions.

Figure 5 shows the distribution of terms with polarity present in the corpus considering positive, negative and neutral terms for both lexicons applied (opLexicon and sentiLex). Considering the terms present in the opLexicon lexicon, it is verified that important terms for the considered domain, such as 'eating' (comida), 'place' (local) and 'American' (americano) are considered neutral, although they mean a lot in the considered context of the restaurant.

[1] Words in Portuguese language.

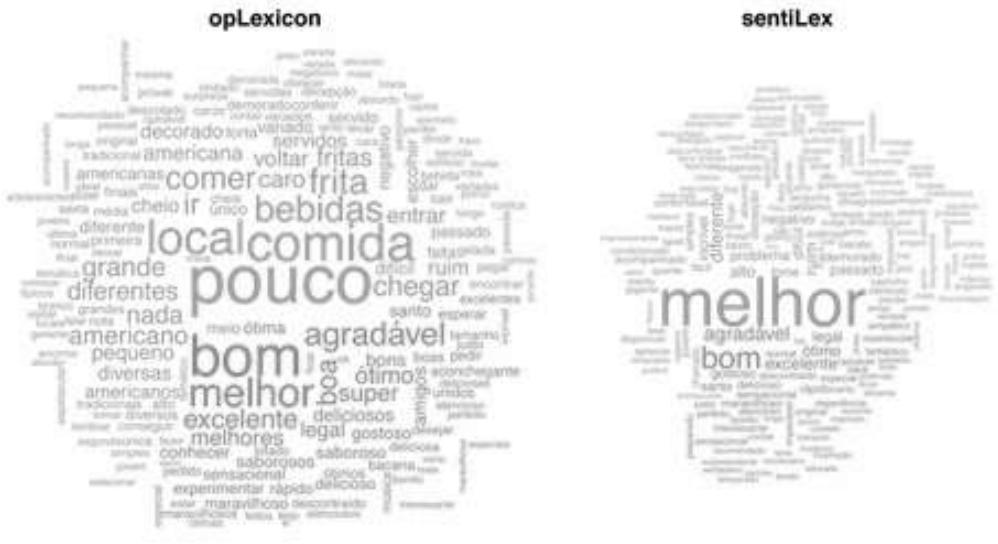

Figure 5 - Cloud of words with positive, negative and neutral terms of COMPANY.
Source: authors.

Another analysis performed by the framework represents the word network from the COMPANY data. For the generation of the word network, the most common bigramas based on the frequency criterion (n> 5) were filtered, as shown in Figure 6.

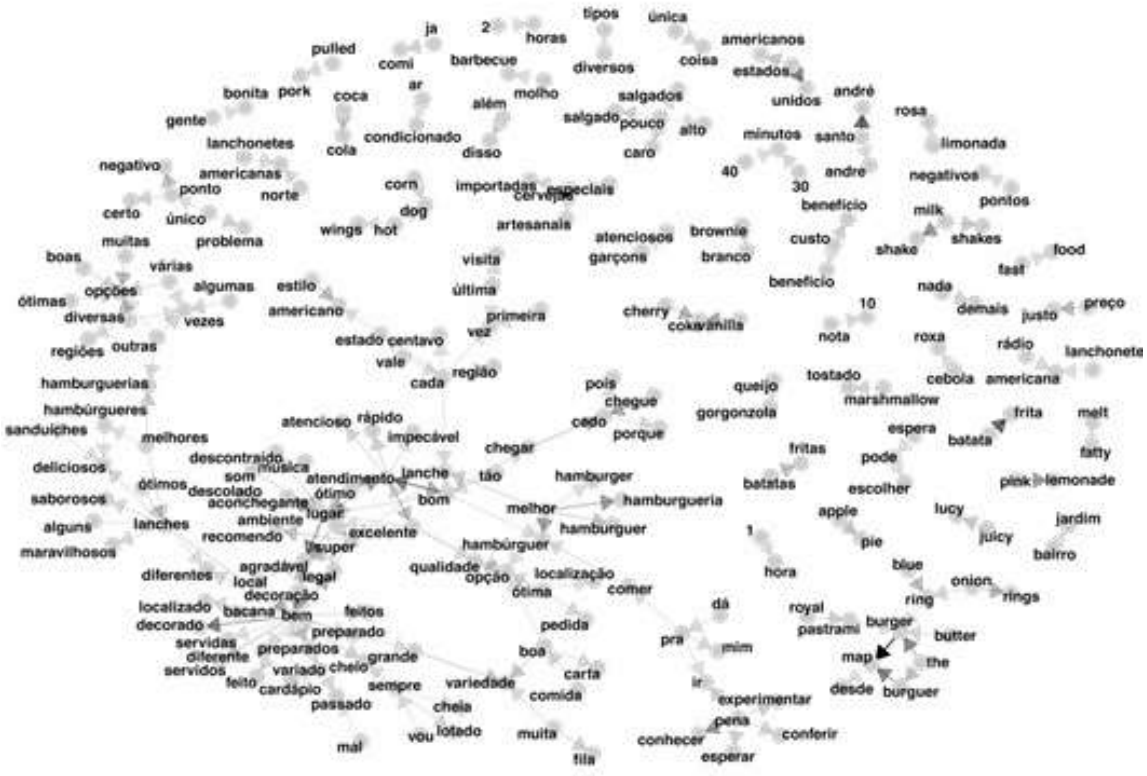

Figure 6 - Graph of relationships between terms generated from COMPANY.

Source: authors.

The network of words shows the relationship between several interesting terms for the domain analyzed in this research, some positive and some negative. Some examples are the relationships between terms like 'muita' (much) and 'fila' (queue) and '2' and 'hours' (horas), such as possibly negative aspects related to waiting time. In counterpoint, we can also see terms like 'worth + each + cent' (‘vale + cada + centavo’) and 'attentive waiters' (‘garçons + atenciosos’), which denote how much people like the food, the place and the service.

In order to finalize the data analysis of COMPANY, the technique of Topic Modeling was applied with the same parameters used so far, which is presented in Figure 7.

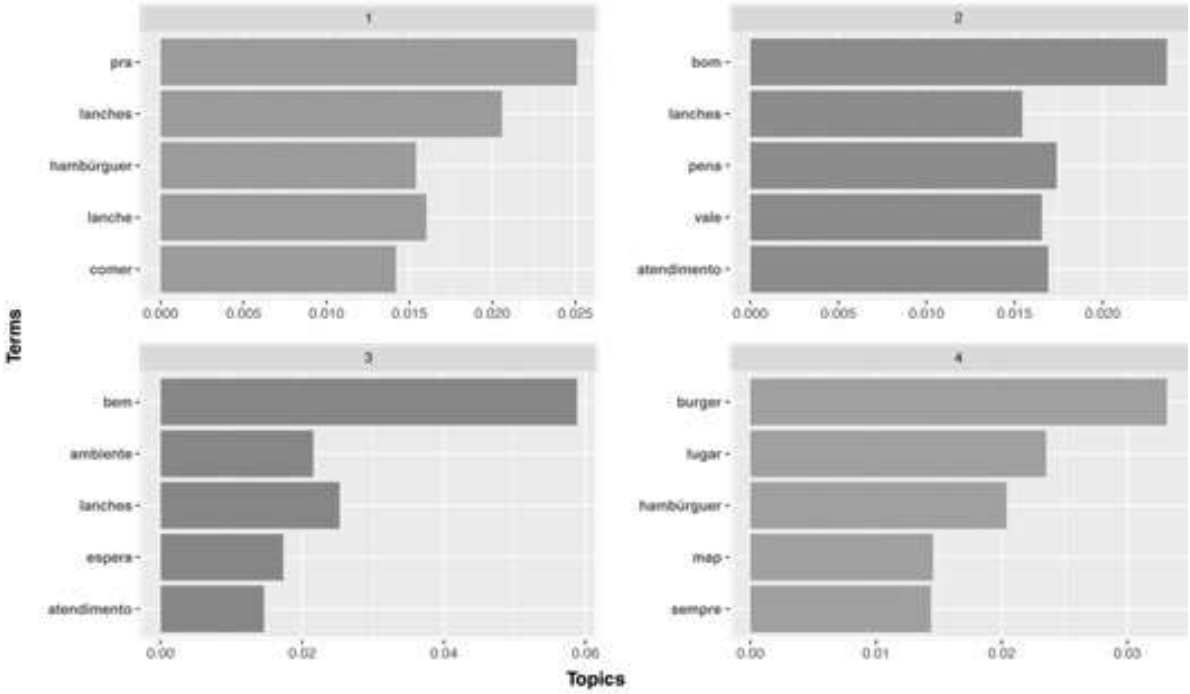

Figure 7 - Topic Modeling for COMPANY data.

Source: authors.

Given the distribution found by the model, it seems that Topic 1 talks about 'food', which is evidenced by the presence of the terms 'snacks' (lanches), 'snack' (lanche) and 'hamburger'. Topic 2 possibly talks about 'experience', given the presence of terms like 'snacks', 'service', 'worth' (‘atendimento’, ‘vale’ and ‘pena’). Topic 3 talks about a mix of topics that have to do with 'service' and 'place', considering the presence of the terms 'service' (atendimento), 'snack' (lanche) and 'wait' (espera). Topic 4 talks about 'place' and 'food', considering the occurrence of terms such as 'place' (lugar), 'burger' and 'hamburger'.

It can be concluded, considering all the analyses carried out, that the customers of COMPANY speak very well about the restaurant. When they think about it, they remember mainly about the food, the place and the service, the three most constant aspects in all the performed analyses.

4. Presentation and discussion of the results

The framework proposed and applied on performed experiment proved to be capable of fulfilling the proposed objectives, especially considering the limitations imposed by the nature of the research, which delimited the universe of data to be processed, in this case a small amount

of data, which in turn has relation with the small size of the companies evaluated in this research.

It has been shown, however, that the Text Mining techniques has the potential to reveal patterns (Feldman & Sanger, 2007) and enables knowledge discover. Thus, the proposed framework can be applied to better understand the customer, his expectations and even his frustrations, thus generating knowledge about customers for the benefit of the company. Such results are aligned with the academic literature, that states that "from a better understanding of the customer, the company will have a better understanding of the true needs and expectations of this customer" (García-Murillo & Annabi, 2002, p. 882).

These analyses revealed, among other indications, that the aspects most addressed by customers refer to food, place and service, varying in intensity and polarity, nonetheless, almost always in a positive way. This may be due to the fact that users' opinions are largely focused on the experience as a whole rather than just one or a few specific aspects.

It is noteworthy in all datasets studied that some of the opinions are quite dense, while others are more succinct, but even so, by applying Text Mining techniques it was possible to have an idea of the frequency with which when certain terms are repeated and how much these influence the final result of the analyses.

Knowing what people talk about most is just the first step. When it comes to relevant information, especially for businesses, it is necessary to know how good or bad people talk when they write about any aspect of their experiences in restaurants. Pang and Lee (2008) affirm that the opinions of other people have always been relevant for people to decide to make decisions about having or not a particular experience. Therefore, the Analysis of Feelings fulfilled an important role for the generation of knowledge about the customer of the analyzed companies. The sentiment analysis conducted in this research was based on lexicons, with the application of two lexicons validated in the Brazilian Portuguese literature: the opLexicon lexicon version 3.0 (Souza & Vieira, 2012) and the sentiLex lexicon (M. J. Silva, Carvalho, Costa, & Sarmento, 2010).

Considering the limitations of the lexicons applied, both presented results that can be considered good, depending on the focus of the analysis. However, it should be considered that neither of the two lexicons was created to be used in the domain addressed in this research (hamburger-type restaurants), which resulted in sometimes peculiar results, as pointed out in several moments of the presented analyses, such as the attribution of negative polarity to

apparently positive terms, considering the nature of the data and the domain considered in this study.

According to Pang and Lee (2008), the sentiment and the subjectivity are very sensitive to the context and to a certain extent dependent on the domain, even considering that the general notion of positive and negative opinions is quite consistent in different domains, as in this research.

In this research, several visualization types were generated in order to provide a more in-depth view of the data. Through the visualization of graphs it was possible to find several connections between terms that appeared in several of the analyses carried out previously, but never arranged as seen in graphs. This leads us to believe that the use of bigrams combined with the visualization of graphs is an excellent way of revealing existing relationships between terms of a corpus, relations that, when approached in other ways, would not provide the same wealth of information.

5. Conclusions

The importance of considering the knowledge coming from the customer for the development of the business is indisputable to successful companies. The contribution of the customers is evidenced through several aspects that directly affect the performance of the companies, however, many of these aspects are subjective and, therefore, related to the tacit knowledge of the customers.

Thus, achieving tacit knowledge of customers is beneficial and essential, especially as a means of ensuring the survival of small and medium-sized enterprises, since this type of organization usually does not have the same level of resources that large companies have.

In this context, Web 2.0 tools, especially Social Networks, play a crucial role, and companies are increasingly dependent on them to interact with their customers, who in turn are increasingly adopting Social Networks as a means of seeking information about companies, products or services and publish opinions about their experiences with these companies.

However, while the number of customer reviews available on social networks makes it a valuable source of information, the task of analyzing this data is not trivial. Considering the increasing volume of these data, it is not a task that can be developed manually by companies, especially those of small and medium size.

In this context, the scenario that motivates this research is: to reconcile techniques of text mining with the objective of revealing knowledge based on the opinions of social network users, especially opinions related to their experiences with restaurants, the domain chosen for the development of this research.

Thus, this research presented a framework for text mining for the discovery of customer knowledge related to their experiences in restaurants from social networks, applicable to the reality of small and medium enterprises.

The main result is the generation of graphical visualizations that contributed to evidence latent relationships between several expressions and terms that were not obvious and that were discovered from the analysis of the applied framework.

The Sentiment Analysis allied to the Topic Modeling revealed that the aspects most addressed by the clients refer to the food, the place and the service, varying in intensity and polarity. In addition, one could infer knowledge about the inclinations of clients' opinions in different contexts.

The literature is rich in terms of applications for Sentiment Analysis. However, most articles that present techniques and tools describe applications in the English language. However, there are not so many papers dedicated to Sentiment Analysis in Portuguese. This research presents an application of Sentiment Analysis in Portuguese that can be considered successful, despite some limitations. In this research Sentiment Analysis is used as a way of inferring knowledge about what clients talk about and how they feel about talking about certain aspects, which corroborates the ideas of Pang and Lee (2008) and Liu (2012), who claim that the simplicity of the method does not undermine its effectiveness.

Another important aspect implemented in this research is the application of topic modeling, through the technique known as Latent Dirichlet Allocation, which in the context of this research fulfills the role of revealing what clients talk when they comment on something related to their experiences.

The results found are consistent with the fundamental idea behind Blei's probabilistic model, although this has been objectively applied, thus generating few topics with only a few items in each one. However, the elaborated model fulfills its design, being able to indicate in the topics addressed, a series of terms that indicate its content, which, in turn, allows us to infer the subject of that topic.

The data used in this experiment were extracted directly from the social network TripAdvisor Brazil, chosen for this research because it contains users reviews about their

experiences in various types of enterprises, among which stand out hotels, restaurants and service providers. For extraction of social network data, we made use of web scraping technique, which consists of direct extraction through a script developed in R language.

The data treatment carried out in this research was carried out in two stages, and was substantially facilitated by the approach adopted, which constitutes in itself a contribution of this work from the technical point of view regarding the organization of the data. These have been treated since their extraction according to the principles of 'tidy data', with the direct consequence of greater flexibility of transformation and treatment.

Regarding the limitations of this research, some factors stand out. The first concerns the universe addressed in this work: Small and Medium Enterprises. Although the methods applied in this research do not apply exclusively to companies of this size, it should be emphasized that some decisions about the construction of the Framework for Opinion Mining applied in this research were influenced by the size of the companies chosen. This was reflected, for example, in the choice of the Sentiment Analysis method, which in the context of this research was based on the use of polarity lexicons to evaluate the inclination of the opinions of the customers of the restaurants whose opinions were analyzed.

Another limitation of this study rightly refers to the use of sentiment polarity lexicons. Considering that there are not many lexicons for the Portuguese language of Brazil, this research applied the two most consistent in the literature in Portuguese.

Finally, another limitation of this research refers to the domain in which it was applied (restaurants). The literature states that there is a certain constancy in relation to the costumers's feeling in different contexts. In this aspect, the use of lexicons that are not thought specifically for the restaurant domain is limited, but they have classifications of terms common to any context.

The results achieved by this research reinforce the importance of investing in ways to understand what customers talk about business in social networks. The applied framework proved to be useful as a tool to better understand the client, his expectations and even his frustrations, allowing to obtain knowledge about the costumers for the benefit of the companies.

As a suggestion of future research, more robust machine learning techniques could be applied due to the increase in the size and complexity of the database. Another suggestion is to apply this framework in different domains. This is a challenge, since, as already mentioned, the domain has a great influence on the results achieved. However, it is believed that by applying

the correct set of techniques, especially involving machine learning, the results can be satisfactory.

References

Ammirato, S., Felicetti, A. M., Gala, M. D., Aramo-Immonen, H., Jussila, J. J., & Kärkkäinen, H. (2019). The use of social media for knowledge acquisition and dissemination in B2B companies: an empirical study of Finnish technology industries. *Knowledge Management Research & Practice*, *17*(1), 52–69. https://doi.org/10.1080/14778238.2018.1541779

Aranha, C. N. (2007). *Processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. Rio de Janeiro: PUC-Rio.

Behringer, N., Sassenberg, K., & Scholl, A. (2017). Knowledge contribution in organizations via social media. *Journal of Personnel Psychology*, 16(1), pp.12-24.

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. Chichester, U.K: Wiley.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Campbell, A. J. (2003). Creating customer knowledge competence: managing customer relationship management programs strategically. *Industrial Marketing Management*, *32*(5), 375–383.

Casey, S. (2017). *The 2016 Nielsen social media report*. Retrieved from The Nielsen Company website: http://www.nielsen.com/us/en/insights/reports/2017/2016-nielsen-social-media-report.html

Chierici, R., Mazzucchelli, A., Garcia-Perez, A., & Vrontis, D. (2019). Transforming big data into knowledge: the role of knowledge management practice. *Management Decision*, 57(8), pp.1902-1922.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Clark, A., Fox, C., & Lappin, S. (Eds.). (2010). *The handbook of computational linguistics and natural language processing*. Chichester, West Sussex; Malden, MA: Wiley-Blackwell.

Correa, T., Hinsley, A. W., & Zúñiga, H. G. de. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, *26*(2), 247–253. https://doi.org/10.1016/j.chb.2009.09.003

Crammond, R., & Murray, A. (2018). Managing knowledge through social media. *Baltic Journal of Management*, 13(3), pp.303-328.

Dalkir, K., & Liebowitz, J. (2011). *Knowledge Management in Theory and Practice* (second edition). Cambridge, Mass: The MIT Press.

Darroch, J., & McNaughton, R. (2003). Beyond market orientation: Knowledge management and the innovativeness of New Zealand firms. *European Journal of Marketing*, *37*(3/4), 572–593. https://doi.org/10.1108/03090560310459096

Davenport, T. H., & Prusak, L. (2000). *Working Knowledge: How Organizations Manage What They Know* (2nd edition). Boston, Mass: Harvard Business Review Press.

Esposito, E., & Evangelista, P. (2016). Knowledge management in SME networks. *Knowledge Management Research & Practice*, *14*(2), 204–212. https://doi.org/10.1057/kmrp.2015.18

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Retrieved from http://www.books24x7.com/marc.asp?bookid=23164

García-Murillo, M., & Annabi, H. (2002). Customer Knowledge Management. *The Journal of the Operational Research Society*, *53*(8), 875–884. https://doi.org/10.1057/palgravejors.2601365

Hoffman, D. L., & Fodor, M. (2010). Can you measure the ROI of your social media marketing? *MIT Sloan Management Review*, *52*(1), 41.

Hofmann, M., & Chisholm, A. (2013). *Text Mining and Visualization - Case Studies Using Open-Source Tools*. Boca Raton, FL: Taylor & Francis Group.

Kaletka, C., & Pelka, B. (2011). Web 2.0 revisited: user-generated content as a social innovation. *International Journal of Innovation and Sustainable Development*, *5*(2–3), 264–275.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Lin, H.-F. (2014). Contextual factors affecting knowledge management diffusion in SMEs. *Industrial Management & Data Systems*, *114*(9), 1415–1437. https://doi.org/10.1108/IMDS-08-2014-0232

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mamorobela, S., & Buckley, S. (2018). Evaluating the effectiveness of social media on knowledge management systems for SMEs. *European Conference on Knowledge Management*, pp.1064-1072

Manivannan, P., & Selvi, C. (2019). Pairwise relative ranking technique for efficient opinion mining using sentiment analysis. *Cluster Computing*, 22(Supplement 6), pp.13487-13497.

Meneghello, J., Thompson, N., Lee, K., Wong, K., & Abu-Salih, B. (2019). Unlocking social media and user generated content as a data source for knowledge management. *Pacrepositoryorg*, 13(1), pp. 1-23.

Nejatian, H., Sentosa, I., Piaralal, S. K., & Bohari, A. M. (2011). The Influence of Customer Knowledge on CRM Performance of Malaysian ICT Companies: A Structural Equation Modeling Approach. *International Journal of Business and Management*, *6*(7). https://doi.org/10.5539/ijbm.v6n7p181

Nisar, T. M., Prabhakar, G., & Strakova, L. (2019). Social media information benefits, knowledge management and smart organizations. *Journal of Business Research*, 94, pp.264-272.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation* (1st ed.). Oxford University Press.

O'reilly, T. (2005). What Is Web 2.0. Retrieved September 5, 2015, from http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Paquette, S. (2011). Customer Knowledge Management. In *Encyclopedia of knowledge management*. Hershey, PA: Idea Group Reference.

Polanyi, M. (1966). *The Tacit Dimension*. Retrieved from https://books.google.com.br/books?id=zfsb-eZHPy0C

Rollins, M., & Halinen, A. (2005). Customer knowledge management competence: Towards a theoretical framework. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 240a–240a. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1385729

Rowley, J. (2002). Eight questions for customer knowledge management in e-business. *Journal of Knowledge Management*, *6*(5), 500–511. https://doi.org/10.1108/13673270210450441

Sebrae. (2014). Participação das micro e pequenas empresas. Retrieved September 20, 2016, from www.sebrae.com.br website: http://www.sebrae.com.br/

Silva, L. A. da, Peres, S. M., & Boscarioli, C. (2017). *Introdução à Mineração de Dados: Com Aplicações em R*. Retrieved from https://books.google.com.br/books?id=5LA4DwAAQBAJ

Silva, M. J., Carvalho, P., Costa, C., & Sarmento, L. (2010). *Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis* (No. TR 10-08). Retrieved from University of Lisbon, Faculty of Sciences, LASIGE website: http://hdl.handle.net/10455/6694

Souza, M., & Vieira, R. (2012). Sentiment Analysis on Twitter Data for Portuguese Language. In H. Caseli, A. Villavicencio, A. Teixeira, & F. Perdigão (Eds.), *Computational Processing of the Portuguese Language* (Vol. 7243, pp. 241–247). https://doi.org/10.1007/978-3-642-28885-2_28

Stewart, T. A. (1997). *Intellectual Capital: The New Wealth of Organizations*. Doubleday / Currency.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.