Knowledge Discovery out of Text Data. A Systematic Review via Text Mining

Journal of Knowledge Management
ISSN: 1367-3270

Publish open access in this journal
Full text online
Content: Table of Contents | Latest Issue RSS RSS
Information: **Journal information** | Editorial Team | Author Guidelines
Other: Journal News (inc. calls for papers) | Sample article | Recommend this journal

Clarivate Analytics 2016 Impact Factor: **2.053***
5-year Impact Factor (2016): **3.293***
Scopus CiteScore 2016: **3.48**
CiteScoreTracker 2017: **3.03** (Updated Month

Search in this title:
Search

Titolo: Knowledge Discovery out of Text Data. A Systematic Review via Text Mining

Authors: Usai, Antonio; Pironti, Marco; Mital, Monika; MEJRI, Chiraz

DOI: 10.1108/JKM-11-2017-0517

Oggetto: Ensuring your article is HEFCE REF-compliant (For UK-based authors only)
Data: venerdì 30 marzo 2018 23:18:34 Ora legale dell'Europa centrale
Da: Journal of Knowledge Management
A: ausai@uniss.it
Allegati: PDF_Proof.PDF

Dear Usai, Antonio; Pironti, Marco; Mital, Monika; MEJRI, Chiraz,

The following email applies to UK-based authors only. If this does not apply to you, please discard this email.

As part of HEFCE's (Higher Education Funding Council for England) REF 2021 (Research Excellence Framework) policy, if you are an author based at a UK research institution and wish to include this article in the next REF assessment, you will need to make a 'closed deposit' of your accepted article to your institutional repository (IR).

A 'closed deposit', as defined by HEFCE, takes place when an Author Accepted Manuscript (AAM) and the surrounding metadata are deposited to the author's institutional repository (IR), with only the article's metadata visible in the IR until the relevant journal embargo period has expired.

Effective from September 2017, Emerald has removed the embargo period across all journals. The full text of the article may therefore become visible within your IR as soon as the final version has been published in the journal.

Note: HEFCE have recently opted to provide greater flexibility on the timing of deposit. For your work to be eligible for REF 2021, you should make this deposit to your IR as soon after the point of acceptance as possible but no later than the points specified in the following HEFCE guidelines:

- Articles accepted for publication between 1 April 2016 and 1 April 2018 should be deposited within three months of publication.

- Articles accepted for publication after 1st April 2018 should be deposited within three months of acceptance for publication.

Please refer to the terms of your own institution to ensure full compliance.

The date your article was accepted for publication was: (30-Mar-2018)

To make your 'closed deposit', you will need to:

• Either, manually deposit the accepted manuscript attached to this email or a version from your own records (it must be the version accepted for publication by the journal's Editor) into your repository
• Or if applicable, forward your accepted manuscript to your institution's Repository Manager.
• If you are unsure which option is applicable to you, please refer to your institutional or departmental publication policy;

In both instances you will need to provide the following information:
• The article DOI (10.1108/JKM-11-2017-0517);
• The article title (''Knowledge Discovery out of Text Data.
A Systematic Review via Text Mining);
• The journal it will be appearing in (Journal of Knowledge Management);
• The article abstract:

The aim of this work is to increase awareness of the potential of the technique of text mining in order to discover knowledge and promote further research collaboration between knowledge management and the information technology communities. From its emergence text mining has involved multidisciplinary studies, focused primarily on database technology, web-based collaborative writing, text analysis, machine learning, and knowledge discovery. However, due to the large amount of research in this field it is becoming more difficult to identify existing studies and therefore suggest new topics.

This article offers a systematic review of 85 academic outputs (articles and books) focused on knowledge discovery derived from the text mining technique. The systematic review is conducted by applying « text mining at the term level, in which knowledge discovery takes place on a more focused collection of words and phrases that are extracted from and label each document » (Feldman et al, 1998, p. 1).
The results revealed the keywords extracted to be associated with the main labels, id est knowledge discovery and text mining can be categorized in two periods: from 1998 to 2009 the term knowledge and text were always used. From 2010 to 2017 in addition to these terms, sentiment analysis, review manipulation, microblogging data, and knowledgeable users were the other terms frequently used. Besides this, it is possible to notice the technical, engineering nature of each term present in the first decade. Whereas, a diverse range of fields such as business, marketing, and finance emerged from 2010 to 2017 due to a greater interest in the online environment.
This is a first comprehensive systematic review on knowledge discovery and text mining through the use of a text mining technique at term level, which offers to reduce redundant research and to avoid the possibility of missing relevant publications.

Please note that the full text of the AAM must only appear in the IR once the final version of the article has been published in the journal.

If you have any questions about Emerald's repository policy and how to make a 'closed deposit', please contact permissions@emeraldinsight.com

For more information on HEFCE's REF 2021 policy, please see:
http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/2016/201635/HEFCE2016_35.pdf.
Thank you for choosing to publish with Emerald, I hope this guidance will help ensure that your work is eligible for REF 2021 inclusion.

Kind Regards,
Chris Tutill
Rights Manager, Emerald Publishing

Knowledge Discovery out of Text Data. A Systematic Review via Text Mining

Journal:	*Journal of Knowledge Management*
Manuscript ID	JKM-11-2017-0517.R2
Manuscript Type:	Research Paper
Keywords:	text mining, data mining, knowledge discovery, data extraction, systematic review, big data analytics

Knowledge Discovery out of Text Data.
A Systematic Review via Text Mining

Abstract

Purpose: The aim of this work is to increase awareness of the potential of the technique of text mining in order to discover knowledge and promote further research collaboration between knowledge management and the information technology communities. From its emergence text mining has involved multidisciplinary studies, focused primarily on database technology, web-based collaborative writing, text analysis, machine learning, and knowledge discovery. However, due to the large amount of research in this field it is becoming more difficult to identify existing studies and therefore suggest new topics.

Design/Methodology: This article offers a systematic review of 85 academic outputs (articles and books) focused on knowledge discovery derived from the text mining technique. The systematic review is conducted by applying « text mining at the term level, in which knowledge discovery takes place on a more focused collection of words and phrases that are extracted from and label each document » (Feldman et al, 1998, p. 1).

Findings: The results revealed the keywords extracted to be associated with the main labels, id est knowledge discovery and text mining can be categorized in two periods: from 1998 to 2009 the term knowledge and text were always used. From 2010 to 2017 in addition to these terms, sentiment analysis, review manipulation, microblogging data, and knowledgeable users were the other terms frequently used. Besides this, it is possible to notice the technical, engineering nature of each term present in the first decade. Whereas, a diverse range of fields such as business, marketing, and finance emerged from 2010 to 2017 due to a greater interest in the online environment.

Originality/Value: This is a first comprehensive systematic review on knowledge discovery and text mining through the use of a text mining technique at term level, which offers to reduce redundant research and to avoid the possibility of missing relevant publications.

Keywords: knowledge discovery, data extraction, big data analytics, text mining, data mining, systematic review

1. Introduction

Before the fourth industrial revolution, in theory and practice KM has been based on the premise that only as human beings are able to draw on the full potential of their brain. In fact, digital organizations are generally notable to fully utilize the knowledge of their employees.

In this line, to achieve such maximum effective usage and so positively influence organizational performance, modern firms seek to acquire or create potentially useful knowledge and to make it available to those who can use it to create value.

It is generally believed if an organization can increase its effective knowledge's utilization by the minimum efforts in sharing and creation, great benefits will come for the firm (King, 2009).

Hence, we are able to state in knowledge intensive firm, organizational learning (OL) as proxy of human being becomes complementary for boosting KM practices in order to turn in routine, information, data and text by embedding into the organization what has been captured from the environment (Levitta and March, 1988).

Since the originality of text mining which emphasizes a new source of knowledge discovering from "hard and intangible documents", we cannot forget it exist a dyadic relationship between knowledge and its stakeholders i.e. people.

In literature, knowledge has been defined as a "justified personal belief" and it appear the fundamental distinction between tacit and explicit knowledge (Polany, 1966).

From one side, we specify tacit knowledge inhabits the minds of people, depending on humans' interpretation and deriving from their knowledge mindset, either difficult to interpret. In a nutshell, knowledge is initially tacit in nature and it will be arduously developed over a long period of time and through trial and error process.

On the other side, we recognize knowledge is often undercapitalizing because "the organization does not know what it knows" (O'Dell and Grayson, 1998).

Since big data analytics based on digital feedbacks coming from the market sets a bridge between information, knowledge, individuals, and organizations that are commonly handled discretely and silently in productive processes and business and marketing activities.

This relationship that have been considered as socially constructed it will be helpful to understand it is impossible to define knowledge universally. Hence, knowledge discovering can only be outlined in practice, during social and digital interactions between individuals.

However, knowledge is embedded the organization and big data we markedly state knowledge discovery are coupled to the individual mind and to contexts. Text mining is considered the process of knowledge discovery by adopting textual databases. It extracts valuable patterns or knowledge from text papers (Balaid, Rozan,Hikmi, & Memon, 2016; Basole, Seuss, & Rouse, 2013; Bookhamer & Zhang, 2016). According to Fan (2006), text data mining is the discovery of new knowledge from "written resources" by using technologies. This has attributed a high commercial value to text mining which generates a new "wave of knowledge discovery" (Tan, 1999). In fact, it is the most natural process of knowledge storage and discovery (Feldman et al., 1998).

Text mining has been used in many domains, such as health care, government, education, and manufacturing. In the text-mining process technologies include pattern matching, topic tracking, summarization, categorization, clustering, association, information visualization (Fan, 2006). Pattern matching is done by analyzing unstructured text and identifying key phrases and relationships within text and it does so by looking for predefined sequences in the text. In turns, it has increased the process of knowledge discovery by employing text mining which has attracted the interest from a variety of fields such as: computer science, business, finance, and artificial intelligence, etc. (Norton, 1999).

This requires more effort to gather non-trivial information but it also provides greater opportunities to implement a knowledge discovery process (KDP). The process allows discerning trivial and non-trivial text data and grasps knowledge present in a large amount of unstructured data. In line with this, Cios et al. (2007) deem that individuals tend to fail in selecting valuable knowledge from text data. The knowledge discovery process calls for knowledge discovery in databases (KDD), based on individuating valid, novel knowledge in data. In a nutshell, the KDP is the means to extract and interpret interesting, non-trivial knowledge, whereas the KDD is the place where knowledge is extracted. The understanding of this process offers several devices on how to manage and shape the interaction between an individual and a machine.

Given this, an increasing number of studies on text mining have been made recently which explore different aspects: Cecchini et al. (2010) present a work on financial words as a predictor of financial events; Cao et al. (2011) investigate the benefits of online review by using a text mining approach; followed by Geva and Zhavi (2014) who argue the empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. And then moving from a customer centered perspective to an internal organizational view, Hogenboom et al. (2016) analyze extraction methods from text for decision support systems.

What has emerged is that text mining has involved multidisciplinary studies, focused on database technology, web-based collaborative writing, text analysis, machine learning, knowledge discovery etc. However, with the increased amount of research in this field it becomes more difficult to identify the existing studies and therefore suggest new topics. For this reason, this article offers a systematic review of academic articles focused on knowledge discovery derived from the text mining technique. The systematic review is conducted by applying « text mining at the term level, in which knowledge discovery takes place on a more focused collection of words and phrases that are extracted from and label each document » (Feldman et al, 1998, p. 1).

The objective is to improve awareness of the potential of this technique in order to discover knowledge and promote further research collaboration between knowledge management and information technology communities.

Literature Review

Due to the phenomenon of globalization, especially in the business to consumer (B2C) market, firms are creating social network driven partnerships and vast amounts of information flow across and within firms, so that more and more businesses are interested in utilizing big data analytics (Reyes & Rosso, 2012; Liu et al., 2014; Lo et al., 2016).

The main focus therefore of scholars on technology and management is to understand the impact of text mining on knowledge discovery in order to aid practitioners to develop big data analytics projects and researchers to conduct new studies (Liu et al., 2014).

In particular, text mining involves working with unstructured or semi-structured data sets such as email, full-text documents, and HTML files. The data are expressed in the form of text which are processed by technologies and converted in structured data (Fan, 2006). According to O'Mara –Eves et al (2015) «text mining is defined as the process of discovering knowledge and structure from unstructured data (i.e. text) (p.2). Currently, texts are extracted also from social media networks such as twitter, blog, facebook etc, to predict market behaviour. In fact review rating and a reviewer's credibility, together with central cues can be used for information searches or for evaluating alternatives (Baek, Ahn, & Choi, 2012; Chung & Tseng, 2012).

Text mining techniques are employed to extract semantic characteristics from review texts. Semantic characteristics are more influential than other characteristics in affecting how many helpfulness votes reviews receive (Cao, Duan, & Gan, 2011).

The decision on which features and classifications can be attributed to a piece of text is crucial because an incorrect attribution of an input will result in a meaningless output. For instance, ''bag-of-words'' is the most commonly used technique for feature selection (Silva et al., 2016). The technique of creating a "bag-of-words'' is to break up a text into words, treat each word as a feature and count the frequency of occurrence of the word and ignore the co-occurrence. Feature selection is the most often used technique among scholars even though the major limitation is that they have ignored the co-occurrence, order of occurrence and association (Schumaker et al., 2012). Alongside, a noun based feature selection is another technique which involves the selection of the noun parts of the speech and then using syntactic rules to find out noun phrases by using used the MUC-7 framework for entity classification. Vu et al. (2012) built a Named Entity Recognition (NER) system for identifying Tweets containing named entities using a linear Conditional Random Field (CRF) model. Syntactic n-grams is another technique which uses a contiguous sequence of n items, i.e. words belonging to a given sequence of text (Butler & Kešelj, 2009; Hagenau, Liebmann, & Neumann, 2013). Additionally, there are a number of combination techniques for short text classification using lexical and semantic features, Bayesian learner, a TF-IDF based selector, a filter-based probabilistic approach, a generative probabilistic model, and Latent Dirichlet Allocation technique (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014).

Besides these, is the technique of feature selection, which is a step in the pre-processing stage of data mining, followed by dimensionality reduction. Therefore dimensionality reduction also becomes critical to the process of text mining (Schumaker & Chen, 2009). Predefined dictionaries can be used for dimensionality reduction such as general use dictionaries like the WordNet thesaurus, or using a term extraction tool to dynamically create a text corpus (Nassirtoussi et al., 2014).

Many studies have used text mining approaches to study the impact of news on market behavior. Schumaker et al. (2012) applies positive and negative sentiment analysis to subjective news articles in order to predict price direction and trading return. Yu et al. (2013) proposed a contextual entropy model to create a thesaurus of emotion words and their corresponding intensities from online stock market news articles. An entropy measure was used to calculate the similarity between the seed words and candidate words and then used to classify the sentiment of the news articles. Hagenau et al. (2013) used a combination of advanced feature extraction methods and a feedback-based feature selection to boost classification accuracy and improve sentiment analytics. According to them, feature selection significantly improves classification accuracies by reducing the number of less-explanatory features, i.e., noise, and thus, may limit negative effects of over-fitting when applying machine learning approaches to classify text messages.

This framework is confirmed primarily with Dorre et al. (1999) that acknowledge that text mining has the same functions of data mining primarily concerning the domain of textual information and relying on sophisticated text analysis techniques that distinguish information from free-text documents. Then, Gupta & Lehal (2009) confirm text mining is the discovery of new, previously unknown information, automatically extracted from different written and computerized resources.

In addition, Tan (1999) proposed a framework which explains the main characteristics of text data mining, referring to the process of extracting interesting and non-trivial patterns or knowledge from text documents. It consist of two components. The first one is related to *text refining* that transforms unstructured text documents into an *intermediate form*. The other is recall as *knowledge distillation* that deduces patterns or knowledge from the *intermediate form*.

Accordingly, the knowledge extraction could be interpreted as: a) vast process with heterogeneous and various data sources, b) independent with dispersed and decentralized control and c) problematic and progressing in data and knowledge associations (Feldman et al., 1998; Mustafa et al., 2009).

Furthermore, Yoon and Park (2004) investigate text-mining networks and provide an high-performance analytical tool in order to analyze technology trends, given standardized and reliable information feedback.

This trend is further confirmed by Mierswa et al. (2006) providing the Yale, a free open source for data extracting from machine learning, able to fuse and form data from multiple sources.

In general, there is a need to carefully design models and metrics that are able to analyze model correlations between disseminated texts and websites and are able to extract the best information from the big data (Ristoski & Paulheim, 2016).

In this regard, literature also discusses the technical challenges related to data samples, structures, heterogeneity of sources, mining models and algorithms, and systems infrastructures that would support data analytics (Khan et al., 2014; Nahm & Mooney, 2012; Larsen & Aone, 1999).

Differently, Li & Lai (2014) argue about the efficiency of mining in smaller samples of data for online purchase:the micro blogosphere, which are in contrast to mining a whole data set, specifically large data sets (e.g. big data). They confirm the advance in analytical applications, specifically when text and data drawing is from large databases (for instance collecting opinions from friends or community of practices). Information is widely interpretative but can create significant opportunities for online sales.

According to the latest developments in this area, the mainstream in literature works on frameworks mapping and predicting learning curves which aim at aiding developed discretionary manipulation proxies in order to study the desired data mining error bound specified by users. Many authors concluded that it is, for the most part, unnecessary to mine a large dataset in volume. Otherwise, it is usually sufficient to sample only up to 1% of the dataset for mining (Jicheng et al., 1999; Cohen & Hersch, 2005; Gophal et al., 2011; Hu et al., 2012). In particular, Lu and Li (2013) have developed an algorithm for bias correction in small samples drawn from online big data (e.g. Twitter and Facebook). They argue that bias mainly depends on the expected number of collisions in the sample.

Contrarily, a stream of researchers acknowledge the lacks of accurate and estimate techniques for consumer reviews in business and marketing analysis, thus introducing an accurate bootstrapping-based framework that estimates results and errors for the different mining techniques (Galitsky et al., 2009; Geva & Zahavi, 2014; Hu et al., 2014).

Finally, several authors focused on the biomedical sector focusing their research activities on reducing bias in specific samples of patients with human disease.

The evaluation of their proposed algorithms shows that they are moderately effective for the feature of knowledge creation from free text applications for studying and practicing the diagnosis and treatment of human disease.

They are considerably more efficient and scalable than some of the existing state-of-the-art batch feature selection algorithms (Uramoto et al., 2004; Cohen & Hersch, 2005; Zhou et al., 2010).

More recently, however, other researchers focused on query optimization techniques over big data in semantic providing repeatable attribution to data scientists when algorithms are not efficient in performing text mining on big data. Thus, they proposed new algorithms based on review texts in order to understand online users' helpfulness voting behaviour for building content-based recommender systems useful to reduce the input/output costs (Ristoski & Paulheim, 2016; Hogenboom et al., 2014; Cao et al., 2011).

Methodology

A systematic review brings results together and shows different perspectives to inform scholars and practitioners to address further research (O'Mara-Eves et al., 2015 : 3). In this line, the authors identify the main publications from 1999 to 2017 showing how the knowledge discovery has been approached by adopting the text mining technique. The scope is to reduce the effect of publication redundancy. Basically, the authors adopt a multi-layered method to searching which is based on a wide Boolean searches of online reference databases, key information, and 'citation trails'.

Additionally, the systematic review is also conducted by applying « text mining at the term level, in which knowledge discovery takes place on a more focused collection of words and phrases that are extracted from and label each document » (Feldman et al, 1998, p. 1). This approach involves extracting labels which correspond to keywords, which consequently represent the main topic of an article. It calls for two approaches : 1). Manual labelling and 2). Automated technique.

The manual labelling is the most difficult but more effective. It requuires an individual who has to label each document in which it can be difficult to extract the right keywords. Whereas, the automated technique tends to be an inaccurate and time-consuming process. This is because it requires a person to label each document first and then the machine can label future documents (Lent et al. 1997).

In this article, the authors used the manual labelling approach, associating to each document with a label derived from « term extraction method». In order to do so, the authors used two main labels which, in taxonomy terms are defined as « knowledge discovery » and « text mining ».

On this basis, a database is made of the year of publication, a topic (which includes part of the abstract or the main information that allows the authors to extract other terms to categorize each paper in one of the main labels, i.e knowledge discovery or text mining), the journal or book, and the author(s). These categories are also known as entities which allow to identify information without having to analyse the primary sources (Norton, 1999). This is in line with Rowley's (1992) statement : « reference databases refer or point the users to another source such as a document, an organization or an individual for additional information or the full text of the document » (p.14).

Therefore, a first screening was made individuating publications where the concepts of knowledge discovery and text mining was reported. In particular, this first screening was composed of two stages. Firstly, we searched in the top 10 conferences and journals in text mining. Therefore we adopted the top five conferences in data mining which are the ACM Conference on Knowledge Discovery and Data Mining, the IEEE International Conference on Data Mining, the International Conference on Information and Knowledge Management (CIKM), the IEEE International Conference on Data Mining (ICDM), and the International Conference on Knowledge Discovering and Data Mining (ACM SIGKDD).

Additionally, we consulted top journals in data mining such as the IEEE Transactions on Knowledge and Data Engineering (TKDE), the Decision Support System, the Technological Forecasting and Social Change, the IITM Journal of Management and IT, Language Learning and Technology and Journal of emerging technologies in web intelligence.

Secondly, we identified top cited papers on Google Scholar, analyzing how the development and deployment of data mining frameworks and methods have also opened the doors for academics and practitioners to knowledge discovery from text (Kayser & Bling, 2017). What emerged was that these concepts were argued primarily from 1998 to 2007. For this period, another screening was done to identify the main topics of knowledge discovery. A table with four columns was created, referring to the aforementioned categories (table 1).

Table 1. First screening

Year	Topic	Journal/Book	Author(s)

1998	Performing a new approach on text mining and knowledge extracting at the term level.	European Symposium on Principles of Data Mining and Knowledge Discovery	Feldman et al.
1999	Present a text mining framework consisting of two components: Text refining and knowledge distillation	Book of Proceedings on Knowledge Discovery from Advanced Databases	Tan
1999	Clustering is a powerful technique for large-scale topic discovery from text, extracting maps each document or record to a point in high-dimensional space	Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining	Larsen & Aone
1999	Application of IBM's Intelligent Miner for Text for analyzing patent portfolios, customer complaint letters, and also competitors' Web pages.	Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining	Dörre et al.
1999	A preliminary analysis discussion about Web mining and information retrieval on the Web by providing a prototype system called WebTMS	IEEE International Conference	Jicheng et al.
2002	Knowledge discovery in text and text mining software	Centre for Research in Information Management	Karanikas & Theodoulidis
2002	A framework for text mining, called DISCOTEX (Discovery from Text Extraction) was introduced by using a learned information extraction system to transform text into more structured data which is then mined for interesting relationships	Proceedings of the AAAI Symposium on Mining Answers from Texts and Knowledge Bases	Nahm & Mooney
2004	Design TAKMI for Biomedical Documents to facilitate knowledge discovery from the very large text databases characteristic of life science and healthcare applications	IEEE Explore	Uramoto et al.
2004	Analytical tool for high-technology trend providing a network-based analysis as alternative method for citation analysis	The Journal of High Technology Management Research	Yoon & Park
2005	A survey on data extraction and text mining on full text access in biomedical literature	Briefings in bioinformatics	Cohen & Hersh
2006	Introduce Yale, a free open-source environment for knowledge discovery database (KDD) and machine learning	Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining	Mierswa et al.
2007	Analyse Data Mining to make sense of a large amount of unsupervised data in some domain by adapting a Discovery Knowledge Driven Approach in order to set a model or knowledge attributes or data	Springer Science & Business Media	Cios et al.

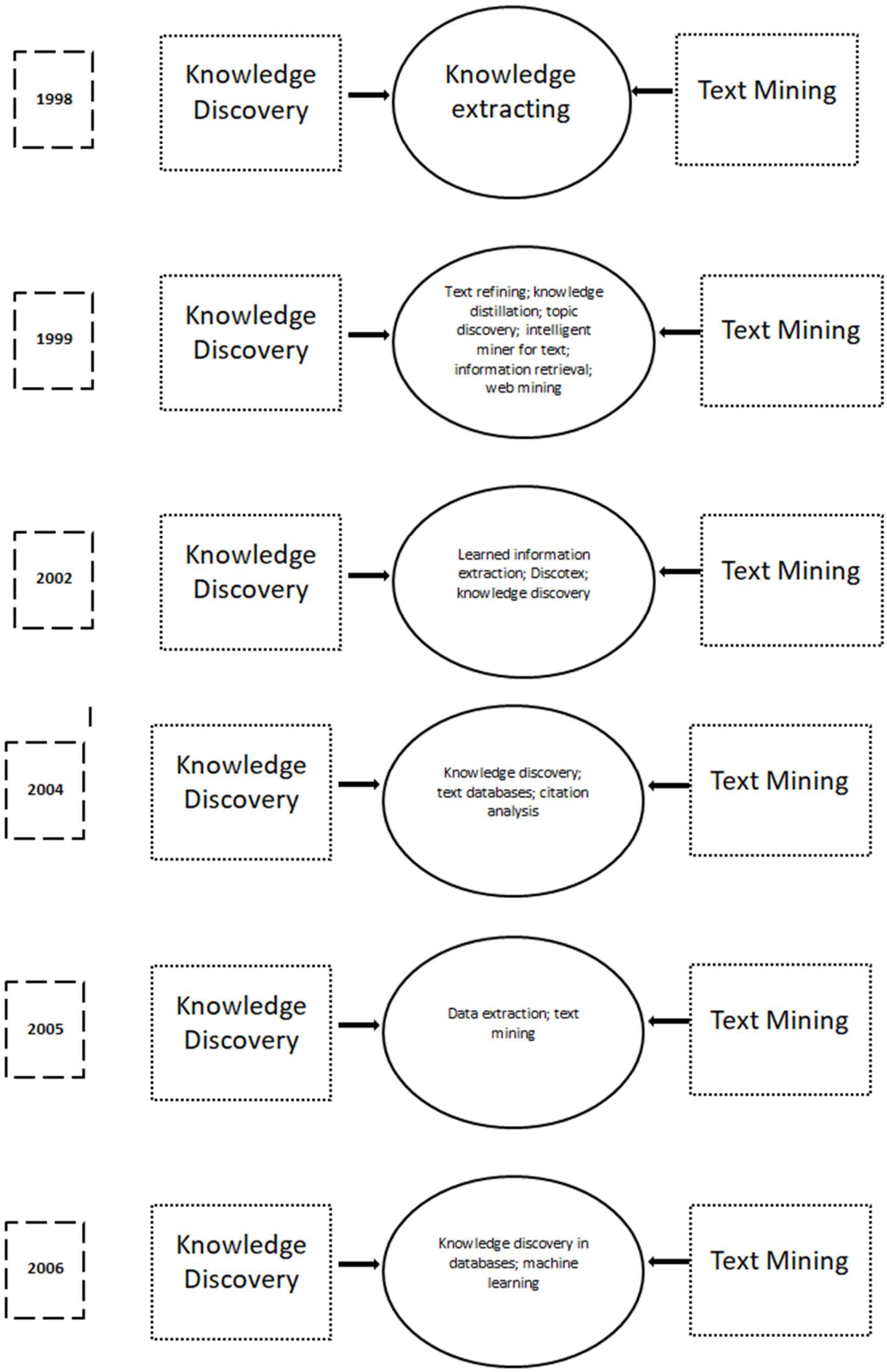
2009	A new approach has been presented for modelling and classifying complaint scenarios associated with customer-firm dialogues, formalized as labelled graphs, in which both firm and customer interact through communicative actions	Decision Support System	Galitsky et al.
2009	Discuss an implementation of Information Extraction and Categorization in the text mining application that we have implemented	CiteSeer	Mustafa et al.
2009	A survey of text mining techniques and applications, by automatically extracting information from different written resources	Journal of emerging technologies in web intelligence	Guptha & Lehal
2010	An algorithm to automatically analyse the emotional polarity of a text and to obtain a value for each piece of text to develop unsupervised text mining approach	Decision Support System	Li & Desheng
2010	Using financial text as a predictor of financial events to aid in discriminating firms that encounter crises for both bankruptcy and fraud merging quantitative data with text data. We achieve our best prediction results	Decision Support System	Cecchini et al.
2010	Extracting meaningful information and knowledge from free text for studying and practicing the diagnosis and treatment of human diseases	Journal of biomedical informatics	Zhou et al.
2010	Sentiment knowledge discovery in twitter streaming data focusing on Twitter data streams pose	International conference on discovery science	Bifet & Frank
2011	Propose a simple statistical method to detect online review manipulation, and assess how consumers respond to products with manipulated reviews	Decision Support System	Hu et al.
2011	Text mining techniques are employed to extract semantic characteristics from review texts in order to understand online users' helpfulness voting behavior	Decision Support System	Cao et al.
2011	Introducing Information Mining exploring how can we transform data into *actionable knowledge*	Decision support System	Gophal et al.
2012	Manipulation of online reviews: An analysis of ratings, readability, and sentiments. The discretionary accrual-based earnings management framework aim at developing a discretionary manipulation proxy to study the management of online reviews.	Decision Support System	Hu et al.
2013	A framework for microblogs was provided	Decision Support System	Li & Li

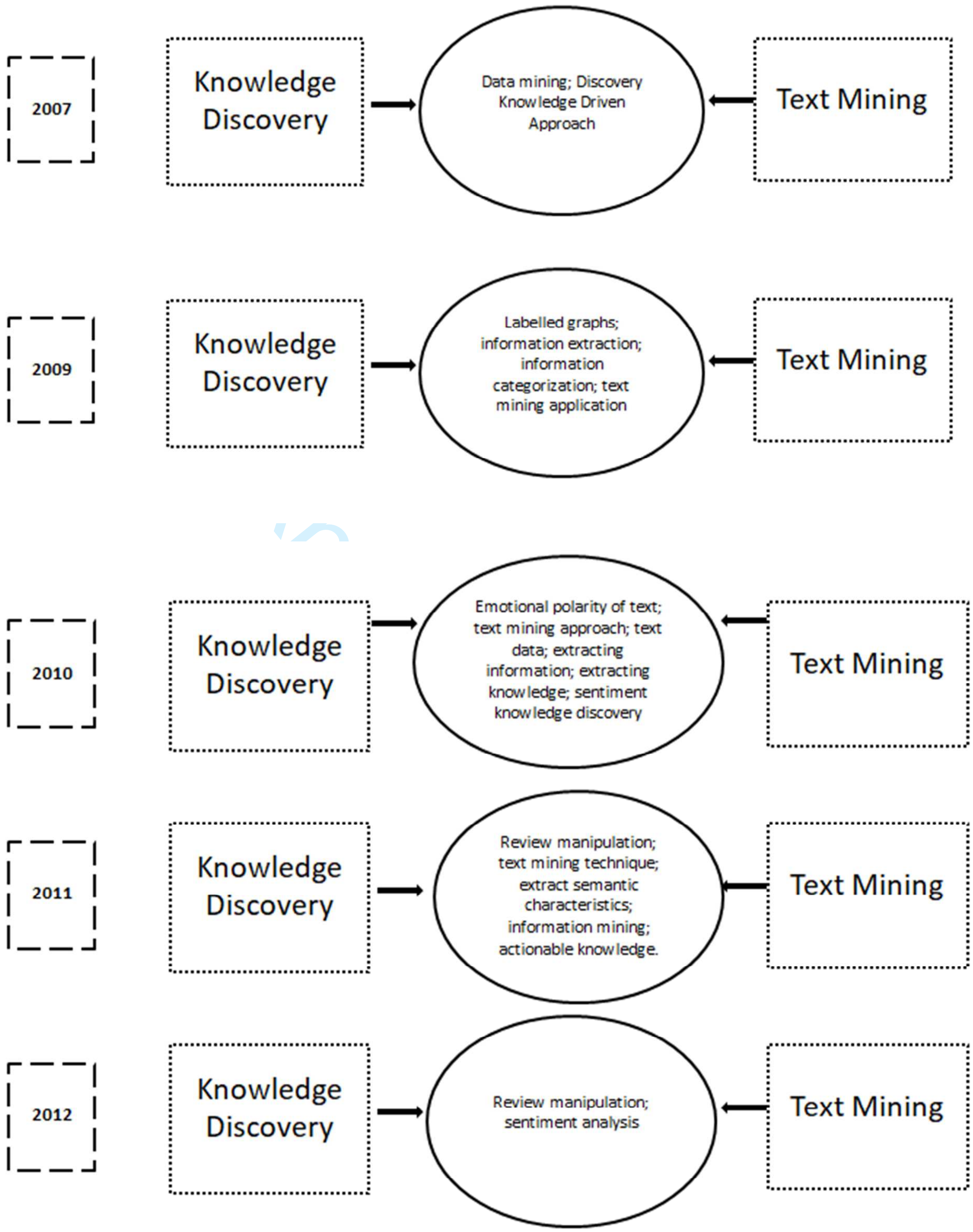
	as compact numeric summarization of opinions from intelligent customers		
2014	Information systems have often been applied to support investors by forecasting price changes in securities markets allowing automated trading engines to appropriately react to news-related liquidity shocks	Decision Support System	Groth et al.
2014	A Multi-lingual support data mining for lexicon-based sentiment analysis guided by semantics improved for the amount of data in different languages on the Web renders.	Decision Support System	Hogemboom et al.
2014	Develop a multiple equation model to examine why the inter-relationships between ratings, sentiments, and sales provide no direct significant accessible and cognitive effort-reducing heuristics in online purchase decisions	Decision Support System	Hu et al.
2014	An Elaboration of an algorithm analysis on sentiment for twitter feed classification of big data fast in order to monitor the publics' feelings towards their brand, business, directors	Decision Support System	Khan et al.
2014	Designing a novel social analytics methodology to analyse sentiments in customer comments that can leverage the sheer volume of consumer reviews archived at social media sites to perform a fine-grained extraction of market intelligence data	Decision Support System	Lau et al.
2014	Empirical evaluation of an automated intraday stock recommendation data incorporating both market data and textual news used to find the incremental value of each knowledge representation, with an end-to-end recommendation process including data pre-processing, modeling, validation, trade recommendations and economic evaluation.	Decision Support System	Geva & Zahavi
2014	Multilevel text mining as a methodology for path knowledge discovery, for linking published research findings in neuropsychiatry	Data Mining and Knowledge Discovery for Big Data	Liu et al.
2014	A methodology of social companionship analysis is proposed for the online users of the micro-blogosphere in order to extract data from participation of knowledgeable users	Decision Support System	Li & Lai
2016	A method called MPM (mining perceptual map) to automatically build perceptual	Decision Support System	Anthony et al.

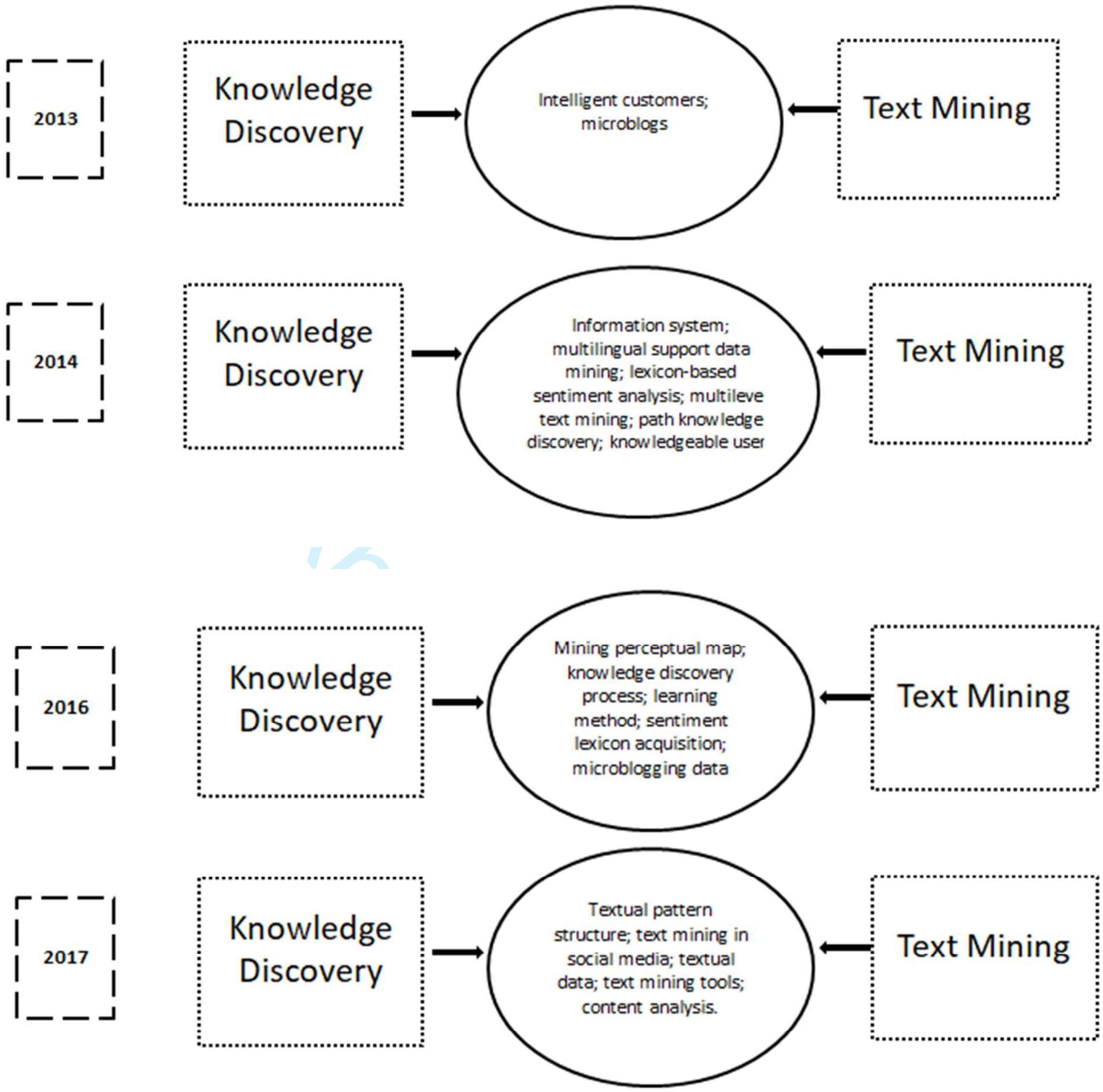
	maps and radar charts from consumer reviews in marketing and business analysis		
2016	A comprehensive overview of those approaches in different stages of the knowledge discovery process for building content-based recommender systems	Web semantics: science, services and agents on the World Wide Web	Ristoski & Paulheim.
2016	A ranking mechanism that is capable of identifying the top-k social audience members using a combination of semi-supervised and supervised learning methods to construct seed words and training data sets with minimal annotation efforts	Decision Support System	Lo et al.
2016	A Stock market sentiment lexicon acquisition has been provided by using microblogging data and statistical measures from three main approaches	Decision Support System	Oliveira et al.
2016	They build a freely available data set with ironic review content that trigger a chain reaction in people and provide valuable knowledge insights from mass and social media market through sentiment analysis, opinion mining and decision making	Decision Support System	Reyes & Rosso
2017	A novel typed textual pattern structure, called Meta PAD	International Conference on Knowledge Discovery and Data Mining	Jiang et al.
2017	The utilization of content data analysin can advise a superior comprehension of the text mining supporters by surveying prescient execution for the expense of outrageous mishaps	International Journal on Reasearch in Computer and Communication Technology	Nagamallika et al.
2017	A survey of text mining in social media: for the purpose of identifying the key themes in the data of Facebook and Twitter	Advances in Science, Technology and Engineering Systems Journal	Salloum et al.
2017	Extending the knowledge base of foresight based on textual data can be accessed and systematically examined through text mining which structures and aggregates data in a largely automated manner	Technological Forecasting and Social Change	Kayser & Bling
2017	They synthesize the current methodological approaches to researching collaborative writing and discuss how new text mining tools can enhance research capacity	Language Learning & Technology	Yim & Warschauer
2017	The paper provides the state of Art on text mining by adapting the Association rules, new techniques of content analysis approach.	IITM Journal of Management and IT	Bhardwaj & Khosla

Following the first part of the analysis, labels\keywords were extracted from the « topic » category as reported in the figure below.

Figure 1. Second screening process


1998
Knowledge Discovery
Knowledge extracting
Text Mining
1999
Knowledge Discovery
Text refining; knowledge distillation; topic discovery; intelligent miner for text; information retrieval; web mining
Text Mining
2002
Knowledge Discovery
Learned information extraction; Discotex; knowledge discovery
Text Mining
2004
Knowledge Discovery
Knowledge discovery; text databases; citation analysis
Text Mining
2005
Knowledge Discovery
Data extraction; text mining
Text Mining
2006
Knowledge Discovery
Knowledge discovery in databases; machine learning
Text Mining


2007
Knowledge Discovery
Data mining; Discovery Knowledge Driven Approach
Text Mining
2009
Knowledge Discovery
Labelled graphs; information extraction; information categorization; text mining application
Text Mining
2010
Knowledge Discovery
Emotional polarity of text; text mining approach; text data; extracting information; extracting knowledge; sentiment knowledge discovery
Text Mining
2011
Knowledge Discovery
Review manipulation; text mining technique; extract semantic characteristics; information mining; actionable knowledge.
Text Mining
2012
Knowledge Discovery
Review manipulation; sentiment analysis
Text Mining



To summarize, we reviewed 85 publications between 1998 and 2017 and then we tried to categorize the results of the text reviewing process, which revealed the papers belonged to two main clusters with both common as well as distinct characteristics. Then, we processed and distributed the articles per year, performing a literature review which sets and predicts the opportunities and challenges for the firms in engaging in social and customer-based data analysis.

Results

It has emerged that from 1998 to 2009 the term knowledge and text were always used. From 2010 to 2017 in addition to these terms, sentiment analysis, review manipulation, microblogging data, and knowledgeable users were the other terms frequently used.

In particular, in 1998 only one publication was found in line with the main labels, i.e knowledge discovery and text mining. In 1999 four papers were anlysed in which emerged : text refining; knowledge distillation; topic discovery; intelligent miner for text; information retrieval; web mining . In 2002 two papers were published using the following terms : Learned information extraction; Discotex; knowledge discovery. In 2003 no papers were found relevant for this study. In 2004 two papers were produced in which knowledge discovery; text databases; and citation analysis were the terms in line with the labels. In 2005 only one article was considered important for this research and data extraction and text mining were the highlighted terms. Again, in 2006 only one paper was considered in line with the labels and knowledge discovery in databases; machine learning were associated to the labels. Followed by a paper published in 2007 in which the data mining and discovery knowledge driven approach were the new concepts used. In 2009, from the analysis of three papers, the key resulted to be: Labelled graphs; information extraction; information categorization; text mining application.

In this first decade, it is possible to note the technical, engineering nature of each term. Whereas, a varied rage of fields emerged from 2010 to 2017, due to a greater interest in the online environment and therefore terms like microblogging, text mining in social media, knowlegeable users, and review manipulation are more used.

In 2010 four papers were considered relevant and the following terms were extracted : Emotional polarity of text; text mining approach; text data; extracting information; extracting knowledge; and sentiment knowledge discovery. In 2011 three papers were analysed in which emerged : review manipulation; text mining technique; extract semantic characteristics; information mining; and actionable knowledge. In 2012 only one paper was considered important where review manipulation and sentiment analysis were the terms used. Again, in 2013 by analysing only onepaper Intelligent customers and microblogs resulted to be the terms in line with the labels. In 2014 eight papers were considered in line with the subjects of this research to which emerged: Information system; multilingual support data mining; lexicon-based sentiment analysis; multilevel text mining; path knowledge discovery; and knowledgeable use. Followed by five articles in 2016 in which the terms extracted were: Mining perceptual map; knowledge discovery process; learning method; sentiment lexicon acquisition; and microblogging data. Concluding with the last six publications where textual pattern structure; text mining in social media; textual data; text mining tools; and content analysis were the terms identified in line with the labels.

Overall, the interest in text mining and knowledge discovery is on the increase exponentially by big data analytics, covering human, consumer, emotional and financial architectures (Li & Desheng, 2010; Cecchini et al., 2010; Jiang et al., 2017; Bhardwaj & Khosla, 2017). The contributions are highly varied and therefore it is very difficult to harvest and share through traditional and manual means.

Discussion and Conclusion

In order to reduce redundant studies and to avoid the possibilities of missing relevant publications, this present research offers a systematic literature review.

More importantly, the interest in data analytics increases exponentially from 2010, especially with the increase of the interest in customer behaviour leading to greater amount of development in the big data arena competition at firm level. Some authors apply sentiment analysis alongside text mining to discover knowledge (Schumaker et al., 2012; Yu et al., 2013). For instance, Hagenau et al. (2013) combines feature extraction methods with sentiment analytics. In turn, websites become the space where disseminated texts are present to extract the best information from the big data (Ristoski & Paulheim, 2016). Hence, texts are analysed in the form of data samples, structures, heterogeneity of sources, mining models and algorithms, and system infrastructures with the aim of generating knowledge (Khan et al., 2014; Nahm & Mooney, 2012). Microbloggings are another foci of interest to evaluate user opinion and therefore create significant opportunities for online sales (Li & Lai, 2014). Small samples drawn from online big data (e.g. Twitter and Facebook) are more investigated to avoid bias (Lu and Li, 2013). This evokes the interest of applying a bootstrapping-based framework that estimates results and errors for the different mining techniques (Geva & Zahavi, 2014; Hu et al., 2014).

Aside from the period of 1998 to 2009, the digital platforms such as twitter, blog, and Facebook, etc. were not so pervasive and therefore less associated with the phenomenon of knowledge discovery and text mining. In fact, text mining was considered having the same functions of data mining Dorre et al. (1999). Although, Tan (1999), defines text mining as the process of extracting interesting and non-trivial patterns by two steps: 1. Text refining and 2. Knowledge distillation. It generates a new knowledge and, as stated by Mustafa et al., (2009), this knowledge is derived of nuances of heterogeneous knowledge and a varied source of data (see also Feldman et al., 1998). This introduces the concept of knowledge discovery more widely. According to Gupta & Lehal (2009), text mining is the discovery of trends extracted from different written and computerized resources. The concept of a multivariate source induces studies on text-mining networks (Yoon and Park, 2004) and free open source for data extracting from machine learning (Mierswa et al., 2006).

In short, in the first period (1998-2009) the studies on knowledge discovery by applying text mining were approached in a more technical way. For instance, there were more studies in the engineering and biomedical sectors. With reference to the latter, the scope was to reduce bias in patients' disease (Uramoto et al., 2004; Cohen & Hersch, 2005). From the second period (2010-2017) more disciplines were interested in this phenomenon. For example, scholars from the business field were developing research to better understand customer behaviour in order to boost sales and reduce costs (Ristoski & Paulheim, 2016; Hogenboom et al., 2014; Cao et al., 2011). As a result, text mining is penetrating various sectors (e.g. medical, e-commerce, health, retail, insurance etc.). This penetration is supported by the overwhelming amount of data available from different sources e.g., web applications, trajectory data, streaming data, RFID, etc. which are growing at an increasing rate. (Chen, Chiang, & Storey, 2012).

According to the results of the text mining analysis, we are able to categorize paper reviews into three main branches depending on common and distinct characteristics. They are: a) technical algorithms, b) framework analysis, c) performance platforms.

This systematic review opens a new research scenario, in contributing to theoretical and practical implications by deriving knowledge and qualitative unstructured patterns from text data set.

From a theoretical point of view, this paper sheds light on past and recent issues, challenges that drive new models and theories in knowledge management studies. In particular, the emphasis on data and text mining represents a new source of complexity for business organizations by validating the tendency of external sources of knowledge (Vrontis et al., 2017). Nowadays, a large amount of information are available in the form of textual data which needs to be categorized and embed in the knowledge management system of a firm (Miao et al., 2009). From a practical point of view, text mining techniques interpret a successful story that will guide the trajectories for consultants, firms and clients in their future environment. Accordingly, practitioners could benefit from research employing the right method for their analysis in order to predict markets. For instance, global economy customers have more sophisticated needs and therefore the urgency of understanding these needs through online text is greatly increased. The online text represents customers' opinions, production quality documentation, and technical knowledge (Kornfein & Goldfrab, 2007). It needs to be converted in operational data to be embraced in the KM process. In turns, it improves product and service quality and offers solution for organizational issues, generating a new knowledge, which can be re-used for further projects. However, knowledge workers and decision makers need to be supported in discovering knowledge patterns (Ur-Rahman and Harding, 2012).

In this line, we recognize there is an increasing interest on text data analytics form both a practical and an academic point of view. In reference to the latter, most prior research has confronted the text of social network analytics but surprisingly there are no insights on how text mining would integrate new communication and marketing processes. Additionally, another gap has individuated on the topic of entrepreneurial knowledge and text mining on social data.

Despite the need for more research, we are also conscious of some limitations related to the present research. First of all, this research is almost exploratory and interpretative and reflects the interpretation of the authors about the framework to use, the theory to understand and text documents to analyze. Therefore, we allow for following this text-mining based approach to analyze literature and, in an attempt to find hidden patterns in the content of documents, the corpus of the study has been text-mined.

Secondly, another limitation is due to the difficulty of understanding patterns to be analysed. Whilst advances in data mining encompass very powerful algorithms, there are a few advances in the literature reviewed on driving the knowledge discovery process towards appropriate results.

In addition, text mining techniques are beginning to encounter problems due to the growing volume of data requiring analysis. Finally, we are aware that new research could be in the form of increased analytics, applied to primary data and according to precise research patterns.

Further research therefore needs to be carried out in order to gather sufficient knowledge regarding this phenomenon. In this way, future directions of research could involve other business sectors by applying specific metrics.

This research tends to aggregate different disciplines, showing the common studies from a wide range of subjects based on online text mining. This is, in fact, a first comprehensive systematic review on knowledge discovery and text mining through the use of a text mining technique at term level, which offers to reduce redundant studies and to avoid the problem of missing relevant publications.

References

Bhardwaj, P., & Khosla, P. (2017). Review of Text Mining Techniques. *IITM Journal of Management and IT*, *8*(1), 27-31.

Bifet, A., & Frank, E. (2010, October). Sentiment knowledge discovery in twitter streaming data. In *International conference on discovery science* (pp. 1-15). Springer, Berlin, Heidelberg.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, *50*(2), 511-521.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*(1), 164-175.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS quarterly*, 1165-1188.

Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data mining: a knowledge discovery approach*. Springer Science & Business Media.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

Dörre, J., Gerstl, P., & Seiffert, R. (1999, August). Text mining: finding nuggets in mountains of textual data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 398-401). ACM.

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., ... & Zamir, O. (1998, September). Text mining at the term level. In *European Symposium on Principles of Data Mining and Knowledge Discovery* (pp. 65-73). Springer, Berlin, Heidelberg.

Galitsky, B. A., González, M. P., & Chesñevar, C. I. (2009). A novel approach for classifying customer complaints through graphs similarities in argumentative dialogues. *Decision Support Systems*, *46*(3), 717-729.

Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision support systems*, *57*, 212-223

Groth, S. S., Siering, M., & Gomber, P. (2014). How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data. *Decision Support Systems*, *62*, 32-42.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

Gopal, J. R. Marsden and J. Vanthienen 2011 Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining. *Decision support systems*, *51*, 4, 721-731.

Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U., & de Jong, F. (2014). Multi-lingual support for lexicon-based sentiment analysis guided by semantics. *Decision support systems*, *62*, 43-53.

Hu, N., Bose, I., Gao, Y., & Liu, L. (2011). Manipulation in digital word-of-mouth: A reality check for book reviews. *Decision Support Systems*, *50*(3), 627-635.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, *52*(3), 674-684.

Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision support systems*, *57*, 42-53.

Jiang, M., Shang, J., Cassidy, T., Ren, X., Kaplan, L. M., Hanratty, T. P., & Han, J. (2017, August). MetaPAD: Meta Pattern Discovery from Massive Text Corpora. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 877-886). ACM.

Jicheng, W., Yuan, H., Gangshan, W., & Fuyan, Z. (1999). Web mining: knowledge discovery on the Web. In *Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on* (Vol. 2, pp. 137-141). IEEE.

Karanikas, H., & Theodoulidis, B. (2002). Knowledge discovery in text and text mining software. *Centre for Research in Information Management, Department of Computation*.

Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. *Technological Forecasting and Social Change*, *116*, 208-215.

Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, *57*, 245-257.

King, W. R. (2009). Knowledge management and organizational learning. In Knowledge management and organizational learning (pp. 3-13). Springer US.

Kornfein, M.M., & Goldfrab, H. (2007). A comparison of classification techniques for technical text passages. In World congress on engineering . London, UK: Proceedings of the World Congress on Engineering.

Larsen, B., & Aone, C. (1999, August). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 16-22). ACM.

Lau, R. Y., Li, C., & Liao, S. S. (2014). Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems*, *65*, 80-94.

Lee, A. J., Yang, F. C., Chen, C. H., Wang, C. S., & Sun, C. Y. (2016). Mining perceptual maps from consumer reviews. *Decision Support Systems*, *82*, 12-25.

Levitt, B., & March, J. G. (1988). Organizational learning. Annual review of sociology, 14(1), 319-338.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems*, *48*(2), 354-368.

Li, Y. M., & Lai, C. Y. (2014). A social appraisal mechanism for online purchase decision support in the micro-blogosphere. *Decision Support Systems*, *59*, 190-205.

Li, Y. M., & Li, T. Y. (2013). Deriving market intelligence from microblogs. *Decision Support Systems*, *55*(1), 206-217.

Liu, C., Chu, W. W., Sabb, F., Parker, D. S., & Bilder, R. (2014). Path knowledge discovery: Multilevel text mining as a methodology for phenomics. In *Data Mining and Knowledge Discovery for Big Data* (pp. 153-192). Springer, Berlin, Heidelberg.

Lo, S. L., Chiong, R., & Cornforth, D. (2016). Ranking of high-value social audiences on Twitter. *Decision Support Systems*, *85*, 34-48.

Miao, D., Duan, Q., Zhang, H., & Jiao, N. (2009). Rough set based hybrid algorithm for text classification. Expert Systems with Applications, 36, 9168–9174.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006, August). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935-940). ACM.

Mustafa, A., Akbar, A., & Sultan, A. (2009). Knowledge discovery using text mining: a programmable implementation on information extraction and categorization.

Nagamallika, G., Anuradha, A., & Sridevi, K. (2017). To Characterize The Contents Of The Documents Through Pattern Discovery In Text Mining. *IJRCCT*, *6*(7), 205-208.

Nahm, U. Y., & Mooney, R. J. (2002, March). Text mining with information extraction. In *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases* (pp. 60-67). Stanford CA.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, *85*, 62-73.

Polanyi, M. (1966). The logic of tacit inference. Philosophy, 41(155), 1-18.

O'dell, C., & Grayson, C. J. (1998). If only we knew what we know: Identification and transfer of internal best practices. California management review, 40(3), 154-174.

Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, *53*(4), 754-760.

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, *36*, 1-22.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), 127-133.

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* (Vol. 8, pp. 65-70).

Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H., & Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal*, *43*(3), 516-533.

Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Systems with Applications, 39(5), 4729-4739.

Vrontis, D., Thrassou, A., Santoro, G., & Papa, A. (2017). Ambidexterity, external knowledge and performance in knowledge-intensive firms. *The Journal of Technology Transfer*, *42*(2), 374-388.

Yim, S., & Warschauer, M. (2017). Web-based collaborative writing in L2 contexts: Methodological insights from text mining. *Language Learning & Technology*, *21*(1), 146-165.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37-50.

Zhou, X., Peng, Y., & Liu, B. (2010). Text mining for traditional Chinese medical knowledge discovery: a survey. *Journal of biomedical informatics*, *43*(4), 650-660.