Knowledge discovery out of text data: a systematic review via text mining

Knowledge discovery out of text data: a systematic review via text mining

Antonio Usai, Marco Pironti, Monika Mital and Chiraz Aouina Mejri

Abstract

Purpose – The aim of this work is to increase awareness of the potential of the technique of text mining to discover knowledge and further promote research collaboration between knowledge management and the information technology communities. Since 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, owing to the large amount of research in this field, it is becoming increasingly difficult to identify existing studies and therefore suggest new topics.

Design/methodology/approach – 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 that 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 owing 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 *Data mining, Systematic review, Text mining, Big data analytics, Knowledge discovery, Data extraction*

1 Introduction

Before the fourth industrial revolution, in theory and practice, knowledge management (KM) has been based on the premise that only human beings are able to draw on the full potential of their brain. In fact, digital organizations are generally not able to fully use their employees' knowledge.

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 make it available to those who can use it to create value.

It is generally believed that if an organization can increase its utilization of effective knowledge by expending minimum efforts in sharing and creation, they will reap great benefits (King, 2009).

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

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

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

From one side, we specify that tacit knowledge inhabits the minds of people, depending on humans' interpretation and deriving from their knowledge mind-set, that is 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 the trial-and-error process.

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

Because big data analytics based on digital feedbacks coming from the market set 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 has been considered as socially constructed will be helpful to understand whether 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 in the organizations, and big data we markedly state knowledge discovery are coupled to the individual mind and to contexts. Text mining is considered as the process of knowledge discovery by adopting textual databases. It extracts valuable patterns or knowledge from text papers (Balaid et al., 2016; Basole et al., 2013; Bookhamer and Zhang, 2016). According to Fan and Li (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. The text mining process, includes technologies such as pattern matching, topic tracking, summarization, categorization, clustering, association, information visualization (Fan and Li, 2006). Pattern matching is done by analyzing the unstructured text and identifying key phrases and relationships within text, and it does so by looking for predefined sequences in the text. In turn, it has increased the process of knowledge discovery by using text mining which has attracted the interest from a variety of fields such as: computer science, business, finance and artificial intelligence, etc.

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 grasping knowledge present in a large amount of unstructured data. In line with this, Cios et al. (2007) deemed that individuals tend to fail in selecting valuable knowledge from text data. KDP calls for knowledge discovery in databases (KDDs), based on individuating valid, novel knowledge in data. In a nutshell, KDP is the means to extract and interpret interesting, non-trivial knowledge, whereas 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, there are an increasing number of studies on text mining recently which explore different aspects: Cecchini et al. (2010) presented a study on financial words as a predictor of financial events; Cao et al. (2011) investigated the benefits of an online review by using the text mining approach; Geva and Zahavi (2014) argued that empirical evaluation of an automated intraday stock recommendation system incorporates both market data and textual news. And then moving from a customer-centered perspective to an internal organizational view, Hogenboom et al. (2014) analyzed extraction methods from text for decision support systems.

What has emerged is that text mining involves multidisciplinary studies and focuses 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 so as to discover knowledge and further promote research collaboration between knowledge management and information technology communities.

2 Literature review

Owing 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 using big data analytics (Reyes and Rosso, 2012; Liu et al., 2014; Lo et al., 2016).

Therefore, scholars' main focus on technology and management is to understand the impact of text mining on knowledge discovery 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 and Li, 2006). According to O'Mara-Eves et al. (2015, p. 2) "text mining is defined as the process of discovering knowledge and structure rom unstructured data" (i.e. text). Currently, texts are also extracted from social media networks such as Twitter, blogs, Facebook, etc., to predict market behavior. In fact review rating and a reviewer's credibility, together with central cues, can be used for information searches or for evaluating alternatives (Baek et al., 2012; Chung and Tseng, 2012).

Text mining techniques are used to extract semantic characteristics from review texts. Semantic characteristics are more influential than other characteristics in affecting how many helpfulness vote reviews are received (Cao et al., 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 (Da Silva et al., 2014). The technique of creating a "bag-of-words" involved breaking up text into words, treating each word as a feature and counting the frequency of occurrence of the word and ignoring 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. Liu 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 and Kesˇelj, 2009; Hagenau et al., 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 Dirichl et al. location technique (Nassirtoussi et al., 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 and 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) applied positive and negative sentiment analysis to subjective news articles 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 Do¨ rre et al. (1999) who acknowledged 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 and Lehal (2009) confirmed 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:

n vast process with heterogeneous and various data sources;

n independent with dispersed and decentralized control; and n 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 a high-performance analytical tool 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 big data (Ristoski and 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 and Mooney, 2002; Larsen and Aone, 1999).

Differently, Li and 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 advancement in analytical applications, specifically

when text and data drawing are 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 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 data set in volume. Otherwise, it is usually sufficient to sample only up to 1 per cent of the data set for mining (Jicheng et al., 1999; Cohen and Hersh, 2005; Gopal 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 portals (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 and 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 and Hersh, 2005; Zhou et al., 2010).

More recently, however, other researchers focused on query optimization techniques over big data in semantics, 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 to understand online users' helpfulness voting behavior for building content-based recommender systems useful to reduce the input/output costs (Ristoski and Paulheim, 2016; Hogenboom et al., 2014; Cao et al., 2011).

3 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, p. 3). In this line, the authors identify the main publications from 1999 to 2017, showing how 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: manual labeling and automated technique.

The manual labelling is the most difficult but more effective. It requires 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 labeling approach, associating each document with a label derived from “term extraction method”. To do so, the authors used two main labels which, in taxonomy terms, are defined as “knowledge discovery” and “text mining”.

Thus, a database 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), is created. These categories are also known as entities which allow to identify information without having to analyze the primary sources. This is in line with 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.”

Therefore, a first screening was made individuating publications where the concepts of knowledge discovery and text mining were reported. In particular, this first screening was composed of two stages. First, we searched in the top ten conferences and journals in text mining. Thereafter, 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.

Second, 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 and Blind, 2017). What emerged was that these concepts were argued primarily from 1998 to 2007. For this period, another screening was

conducted to identify the main topics of knowledge discovery. A table with four columns was created, referring to the aforementioned categories (Table I).

Following the first part of the analysis, labels\keywords were extracted from the "topic" category as reported in Figure 1.

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 and 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.

4 Results

It has emerged that from 1998 to 2009, the terms 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 analyzed, in which text refining, knowledge distillation, topic discovery, intelligent miner for text, information retrieval and Web mining emerged. In 2002, two papers were published using the following terms: learned information extraction and DiscoTex – knowledge discovery. In 2003 no relevant papers were found 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 KDD; machine learning was associated with 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: labeled graphs, information extraction, information categorization and text mining application.

In this first decade, it is possible to note the technical, engineering nature of each term. Whereas, a varied range of fields emerged from 2010 to 2017, owing to a greater interest in the online environment, and therefore, terms like microblogging, text mining in social media, knowledgeable 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 analyzed in which review manipulation, text mining technique, extract semantic characteristics, information mining and actionable knowledge emerged. In 2012 only one paper was considered important where review manipulation and sentiment analysis were the, terms used. Again, in 2013, by analyzing only one paper, 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 information system, multilingual support data mining, lexicon-based sentiment analysis, multilevel text mining, path knowledge discovery and knowledgeable use emerged. Followed by five articles in 2016 in which the terms extracted were: mining perceptual map, KDP, 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 and Wu, 2010; Cecchini et al., 2010; Jiang et al., 2017; Bhardwaj and Khosla, 2017).

The contributions are highly varied, and therefore, it is very difficult to harvest and share through traditional and manual means.

5 Discussion and conclusion

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 behavior, 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) combined feature extraction methods with sentiment analytics. In turn, websites become the space where disseminated texts are present to extract the best information from big data (Ristoski and Paulheim, 2016). Hence, texts are analyzed 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 and Mooney, 2002). Microblogs are another foci of interest to evaluate user opinion and therefore create significant opportunities for online sales (Li and Lai, 2014). Small samples drawn from online big data portals (e.g. Twitter and Facebook) are investigated in depth 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 and Zahavi, 2014; Hu et al., 2014).

Aside from the period from 1998 to 2009, digital platforms such as Twitter, blogs and Facebook, etc. were not so pervasive and therefore less associated with the phenomena of knowledge discovery and text mining. In fact, text mining was considered having the same functions as data mining (Dörre et al., 1999). However, Tan (1999) defined text mining as the process of extracting interesting and non-trivial patterns by two steps: text refining and knowledge distillation. It generates new knowledge and, as stated by Mustafa et al. (2009), this knowledge is derived from nuances of heterogeneous knowledge and a varied source of data (Feldman et al., 1998). This introduces the concept of knowledge discovery more widely. According to Gupta and 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 and Hersh, 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 behavior to boost sales and reduce costs (Ristoski and 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 et al., 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: technical algorithms, framework analysis and performance platforms.

This systematic review opens a new research scenario toward 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 is 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 by using the right methods for their analysis to predict markets. For instance, global economy customers have more sophisticated needs and therefore the urgency for understanding these needs through online text is greatly increased. The online text represents customers' opinions, production quality documentation and technical knowledge (Kornfein and Goldfrab, 2007). It needs to be converted in operational data to be embraced in the KM process. In turn, 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 have 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, 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 follow 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.

Second, another limitation is the difficulty of understanding patterns to be analyzed. While advances in data mining encompass very powerful algorithms, there are a few advances in the literature reviewed on driving the KDP toward appropriate results.

In addition, text mining techniques are beginning to encounter problems because of 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 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

Baek, H., Ahn, J. and Choi, Y. (2012), “Helpfulness of online consumer reviews: readers’ objectives and review cues”, International Journal of Electronic Commerce, Vol. 17 No. 2, pp. 99-126.

Balaid, A., Rozan, M.Z.A., Hikmi, S.N. and Memon, J. (2016), “Knowledge maps: a systematic literature review and directions for future research”, International Journal of Information Management, Vol. 36 No. 3, pp. 451-475.

Basole, R.C., Seuss, C.D. and Rouse, W.B. (2013), “IT innovation adoption by enterprises: knowledge discovery through text analytics”, Decision Support Systems, Vol. 54 No. 2, pp. 1044-1054.

Bhardwaj, P. and Khosla, P. (2017), “Review of text mining techniques”, IITM Journal of Management and IT, Vol. 8 No. 1, pp. 27-31.

Bifet, A. and Frank, E. (2010), “Sentiment knowledge discovery in twitter streaming data”, International Conference on Discovery Science, Springer, Berlin, Heidelberg, pp. 1-15.

Bookhamer, P. and Zhang, Z.J. (2016), “Knowledge management in a global context: a case study”, Information Resources Management Journal (IRMJ), Vol. 29 No. 1, pp. 57-74.

Butler, M. and Kesˇelj, V. (2009), "Financial forecasting using character n-gram analysis and readability scores of annual reports", Canadian Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, pp. 39-51.

Cao, Q., Duan, W. and Gan, Q. (2011), "Exploring determinants of voting for the 'helpfulness' of online user reviews: a text mining approach", Decision Support Systems, Vol. 50 No. 2, pp. 511-521.

Cecchini, M., Aytug, H., Koehler, G.J. and Pathak, P. (2010), "Making words work: using financial text as a predictor of financial events", Decision Support Systems, Vol. 50 No. 1, pp. 164-175.

Chen, H., Chiang, R.H. and Storey, V.C. (2012), "Business intelligence and analytics: from big data to big impact", MIS Quarterly, Vol. 35 No. 4, pp. 1165-1188.

Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L.A. (2007), Data Mining: A Knowledge Discovery Approach, Springer Science & Business Media, Berlin.

Cohen, A.M. and Hersh, W.R. (2005), "A survey of current work in biomedical text mining", Briefings in Bioinformatics, Vol. 6 No. 1, pp. 57-71.

Da Silva, N.F., Hruschka, E.R. and Hruschka, E.R. Jr (2014), "Tweet sentiment analysis with classifier ensembles", Decision Support Systems, Vol. 66, pp. 170-179.

Dörre, J., Gerstl, P. and Seiffert, R. (1999), "Text mining: finding nuggets in mountains of textual data", Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 398-401.

Fan, J. and Li, R. (2006), "Statistical challenges with high dimensionality: feature selection in knowledge discovery", arXiv preprint math/0602133.

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

Galitsky, B.A., Gonza´lez, M.P. and Chesn˜evar, C.I. (2009), "A novel approach for classifying customer complaints through graphs similarities in argumentative dialogues", Decision Support Systems, Vol. 46 No. 3, pp. 717-729.

Geva, T. and Zahavi, J. (2014), "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news", Decision Support Systems, Vol. 57 No. 1, pp. 212-223.

Gopal, J., Marsden, R. and Vanthienen, J. (2011), "Information mining – reflections on recent advancements and the road ahead in data, text, and media mining", Decision Support Systems, Vol. 51 No. 4, pp. 721-731.

Groth, C.W., Wimmer, M., Akhmerov, A.R. and Waintal, X. (2014), "Kwant: a software package for quantum transport", New Journal of Physics, Vol. 16 No. 6, p. 63.

Gupta, V. and Lehal, G.S. (2009), "A survey of text mining techniques and applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1 No. 1, pp. 60-76.

Hagenau, M., Liebmann, M. and Neumann, D. (2013), "Automated news reading: stock price

prediction based on financial news using context-capturing features", Decision Support Systems, Vol. 55 No. 3, pp. 685-697.

Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U. and de Jong, F. (2014), "Multi-lingual support for lexicon-based sentiment analysis guided by semantics", Decision Support Systems, Vol. 62 No. 6, pp. 43-53.

Hu, N., Koh, N.S. and 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, Vol. 57 No. 1, pp. 42-53.

Hu, N., Bose, I., Gao, Y. and Liu, L. (2011), "Manipulation in digital word-of-mouth: a reality check for book reviews", Decision Support Systems, Vol. 50 No. 3, pp. 627-635.

Hu, N., Bose, I., Koh, N.S. and Liu, L. (2012), "Manipulation of online reviews: an analysis of ratings, readability, and sentiments", Decision Support Systems, Vol. 52 No. 3, pp. 674-684.

Jiang, M., Shang, J., Cassidy, T., Ren, X., Kaplan, L.M., Hanratty, T.P. and Han, J. (2017), "MetaPAD: meta pattern discovery from massive text corpora", Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 877-886.

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

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

Kayser, V. and Blind, K. (2017), "Extending the knowledge base of foresight: the contribution of text mining", Technological Forecasting and Social Change, Vol. 116 No. 3, pp. 208-215.

Khan, F.H., Bashir, S. and Qamar, U. (2014), "TOM: Twitter opinion mining framework using hybrid classification scheme", Decision Support Systems, Vol. 57 No. 1, pp. 245-257.

King, W.R. (2009), "Knowledge management and organizational learning", Knowledge Management and Organizational Learning, Springer, pp. 3-13.

Kornfein, M.M. and Goldfrab, H. (2007), "A comparison of classification techniques for technical text passages", Proceedings of the World Congress on Engineering, World Congress on Engineering, London.

Larsen, B. and Aone, C. (1999), "Fast and effective text mining using linear-time document clustering", Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 16-22.

Lau, R.Y., Li, C. and Liao, S.S. (2014), "Social analytics: learning fuzzy product ontologies for aspect-oriented sentiment analysis", Decision Support Systems, Vol. 65 No. 9, pp. 80-94.

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

Lent, B., Agrawal, R. and Srikant, R. (1997), "Discovering trends in text databases", KDD, Vol. 97, pp.

227-230.

Levitt, B. and March, J.G. (1988), “Organizational learning”, Annual Review of Sociology, Vol. 14 No. 1, pp. 319-338.

Li, N. and Wu, D.D. (2010), “Using text mining and sentiment analysis for online forums hotspot detection and forecast”, Decision Support Systems, Vol. 48 No. 2, pp. 354-368.

Li, Y.M. and Lai, C.Y. (2014), “A social appraisal mechanism for online purchase decision support in the micro-blogosphere”, Decision Support Systems, Vol. 59 No. 3, pp. 190-205.

Li, Y.M. and Li, T.Y. (2013), “Deriving market intelligence from microblogs”, Decision Support Systems, Vol. 55 No. 1, pp. 206-217.

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

Liu, X., Zhou, M., Wei, F., Fu, Z. and Zhou, X. (2012), “Joint inference of named entity recognition and normalization for tweets”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Long Papers-Volume 1, Association for Computational Linguistics, pp. 526-535.

Lo, S.L., Chiong, R. and Cornforth, D. (2016), “Ranking of high-value social audiences on twitter”,Decision Support Systems, Vol. 85, pp. 34-48.

Lu, J. and Li, D. (2013), “Bias correction in a small sample from big data”, IEEE Transactions on Knowledge and Data Engineering, Vol. 25 No. 11, pp. 2658-2663.

Miao, D., Duan, Q., Zhang, H. and Jiao, N. (2009), “Rough set based hybrid algorithm for text classification”, Expert Systems with Applications, Vol. 36 No. 5, pp. 9168-9174.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. (2006), “Yale: rapid prototyping for complex data mining tasks”, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 935-940.

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

Nagamallika, G., Anuradha, A. and Sridevi, K. (2017), “To characterize the contents of the documents through pattern discovery in text mining”, IJRCCT, Vol. 6 No. 7, pp. 205-208.

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

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y. and Ngo, D.C.L. (2014), “Text mining for market prediction: a systematic review”, Expert Systems with Applications, Vol. 41 No. 16, pp. 7653-7670.

O’dell, C. and Grayson, C.J. (1998), “If only we knew what we know: identification and transfer of internal best practices”, California Management Review, Vol. 40 No. 3, pp. 154-174.

Oliveira, N., Cortez, P. and Areal, N. (2016), “Stock market sentiment lexicon acquisition using microblogging data and statistical measures”, Decision Support Systems, Vol. 85 No. 5, pp. 62-73.

O’Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Ananiadou, S. (2015), “Using text mining for study identification in systematic reviews: a systematic review of current approaches”, Systematic Reviews, Vol. 4 No. 1, p. 5.

Polanyi, M. (1966), “The logic of tacit inference”, Philosophy, Vol. 41 No. 155, pp. 1-18.

Reyes, A. and Rosso, P. (2012), “Making objective decisions from subjective data: detecting irony in customer reviews”, Decision Support Systems, Vol. 53 No. 4, pp. 754-760.

Ristoski, P. and 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, Vol. 36 No. 1, pp. 1-22.

Salloum, S.A., Al-Emran, M., Monem, A.A. and Shaalan, K. (2017), “A survey of text mining in social media: Facebook and twitter perspectives”, Advances in Science, Technology and Engineering Systems Journal, Vol. 2 No. 1, pp. 127-133.

Schumaker, R.P. and Chen, H. (2009), “Textual analysis of stock market prediction using breaking financial news: the AZFin text system”, ACM Transactions on Information Systems (TOIS), Vol. 27 No. 2, p. 12.

Schumaker, R.P., Zhang, Y., Huang, C.N. and Chen, H. (2012), “Evaluating sentiment in financial news articles”, Decision Support Systems, Vol. 53 No. 3, pp. 458-464.

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

Ur-Rahman, N. and Harding, J.A. (2012), “Textual data mining for industrial knowledge management and text classification: a business oriented approach”, Expert Systems with Applications, Vol. 39 No. 5, pp. 4729-4739.

Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. and Takeda, K. (2004), “A text-mining system forknowledge discovery from biomedical documents”, IBMSystems Journal, Vol. 43 No. 3, pp. 516-533.

Vrontis, D., Thrassou, A., Santoro, G. and Papa, A. (2017), “Ambidexterity, external knowledge and performance in knowledge-intensive firms”, The Journal of Technology Transfer, Vol. 42 No. 2, pp. 374-388.

Yim, S. and Warschauer, M. (2017), “Web-based collaborative writing in L2 contexts: methodological insights from text mining”, Language Learning & Technology, Vol. 21 No. 1, pp. 146-165.

Yoon, B. and Park, Y. (2004), “A text-mining-based patent network: analytical tool for high-technology trend”, The Journal of High Technology Management Research, Vol. 15 No. 1, pp. 37-50.

Yu, L.C., Wu, J.L., Chang, P.C. and Chu, H.S. (2013), “Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news”,

Knowledge- Based Systems, Vol. 41, pp. 89-97.

Zhou, X., Peng, Y. and Liu, B. (2010), “Text mining for traditional Chinese medical knowledge discovery: a survey”, Journal of Biomedical Informatics, Vol. 43 No. 4, pp. 650-660.

Further reading

Chen, Y.T. and Chou, T.Y. (2012), “Exploring the continuance intentions of consumers for B2C online shopping: perspectives of fairness and trust”, Online Information Review, Vol. 36 No. 1, pp. 104-125.

Lee, A.J., Yang, F.C., Chen, C.H., Wang, C.S. and Sun, C.Y. (2016), “Mining perceptual maps from consumer reviews”, Decision Support Systems, Vol. 82 No. 2, pp. 12-25.

Li, N. and Wu, D.D. (2010), “Using text mining and sentiment analysis for online forums hotspot detection and forecast”, Decision Support Systems, Vol. 48 No. 2, pp. 354-368.

Sun, B.Y., Li, J., Wu, D.D., Zhang, X.M. and Li, W.B. (2010), “Kernel discriminant learning for ordinal regression”, IEEE Transactions on Knowledge and Data Engineering, Vol. 22 No. 6, pp. 906-910.

Author(s)/ Year	*Topic*	*Journal/Book*
Feldman et al. (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
Tan (1999	Present a text mining framework consisting of two components: text refining and knowledge distillation	Book of Proceedings on Knowledge Discovery from Advanced Databases
Larsen and Aone (1999)	Clustering is a powerful technique for largescale topic discovery from text, extracting maps each document or record to a point in high-dimensional space	Proceedings of ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining
Dörre et al.(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
Jicheng et al. (1999) IEEE International Conference	A preliminary analysis discussion about Web mining and information retrieval on the Web by providing a prototype system called WebTMS	IEEE International Conference
Karanikas and Theodoulidis (2002)	Knowledge discovery in text and text mining software	Centre for Research in InformationManagement
Nahm and Mooney (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
Uramoto et al. (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
Yoon and Park (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
Cohen and Hersh (2005)	A survey on data extraction and text mining on full text access in biomedical literature	Briefings in Bioinformatics
Mierswa et al.(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
Cios 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
Galitsky 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
Mustafa et al. (2009)	Discuss an implementation of information extraction and categorization in the text mining application that we have implemented	CiteSeer

Gupta andLehal (2009)	A survey of text mining techniques and applications, by automatically extracting information from different written resources	Journal of Emerging Technologies in Web Intelligence
Li and Wu (2010)	An algorithm to automatically analyze 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
Cecchini et al. (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
Zhou 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
Bifet and Frank (2010)	Sentiment knowledge discovery in twitter streaming data focusing on Twitter data streams pose	International Conference on Discovery Science
Hu et al. (2011)	Propose a simple statistical method to detect online review manipulation, and assess how consumers respond to products with manipulated reviews	Decision Support System
Cao et al.(2011)	Text mining techniques are employed to extract semantic characteristics from review texts in order to understand online users' helpfulness voting behaviour	Decision Support System
Gopal et al. (2011)	Introducing Information Mining exploringhow can we transform data into actionable knowledge	Decision Support System
Hu 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
Li and Li (2013)	A framework for microblogs was provided as compact numeric summarization of opinions from intelligent customers	Decision Support System
Groth et al. (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
Hogenboom 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
Hu 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
Khan 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
Lau et al. (2014)	Designing a novel social analytics methodology to analyze sentiments in customer comments that can leverage the sheer volume of consumer reviews archived at social media sites to perform a finegrained extraction of market intelligence data	Decision Support System

Geva and Zahavi (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
Liu et al. (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
Li and Lai (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
Lee et al. (2016)	A method called MPM (mining perceptual map) to automatically build perceptual maps and radar charts from consumer reviews in marketing and business analysis	Decision Support System
Ristoski and Paulheim (2016)	A comprehensive overview of those approaches in different stages of the KDP for building content-based recommender systems	Web semantics: Science, Services and Agents on the World Wide Web
Lo et al. (2016)	A ranking mechanism that is capable of identifying the top-k social audience members using a combination of semisupervised and supervised learning methods to construct seed words and training data sets with minimal annotation efforts	Decision Support System
Oliveira 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
Reyes and Rosso (2012)	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
Jiang et al.(2017)	A novel typed textual pattern structure, called Meta PAD	International Conference on Knowledge Discovery and Data Mining
Nagamallika et al. (2017)	The utilization of content data analysis can advise a superior comprehension of the text mining supporters by surveying prescient execution for the expense of outrageous mishaps	International Journal on Research in Computer and Communication Technology
Salloum 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
Kayser and Bling (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
Yim and Warschauer (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

Bhardwaj and Khosla (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

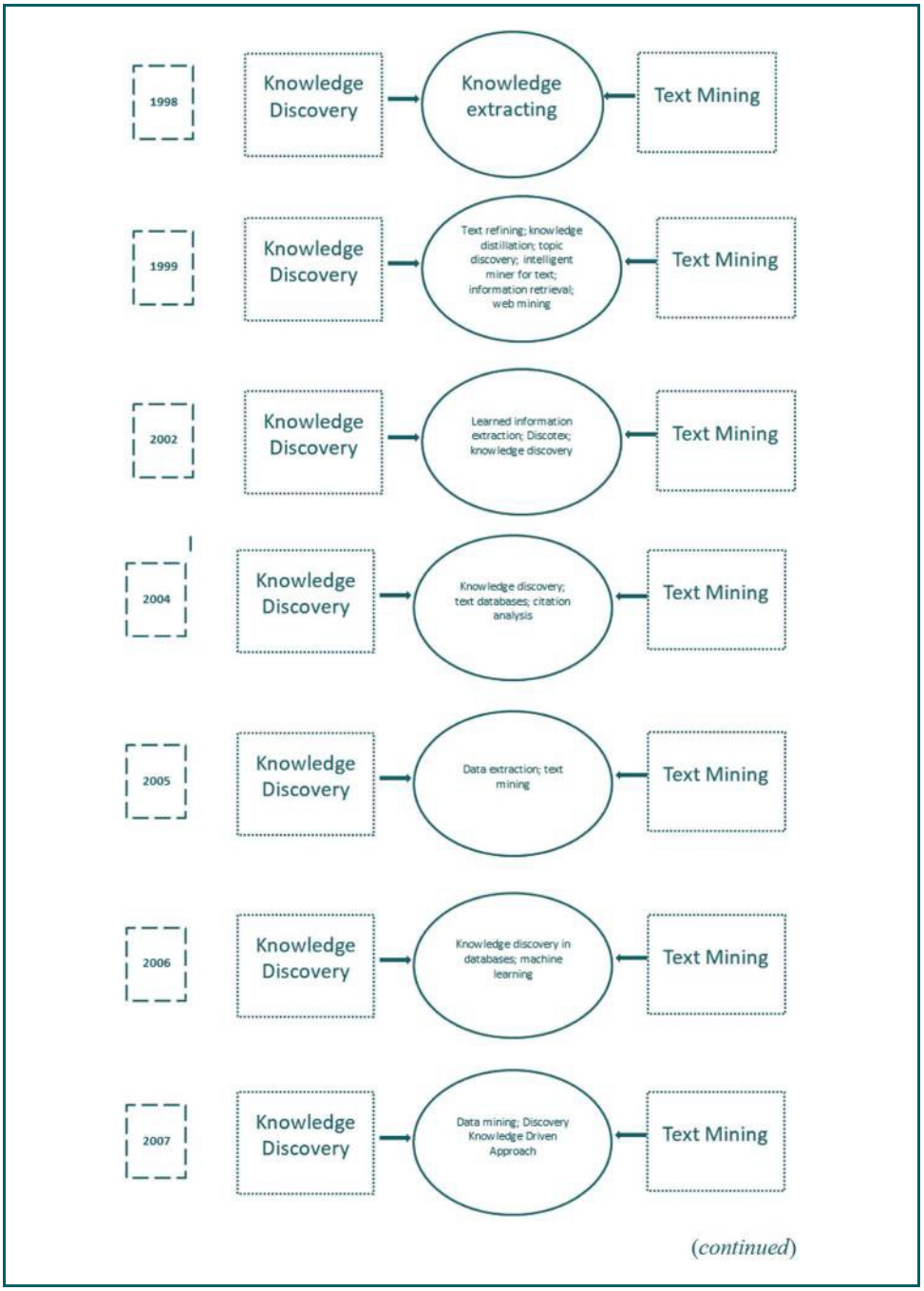

Figure 1 Second Screening Progress (continued)

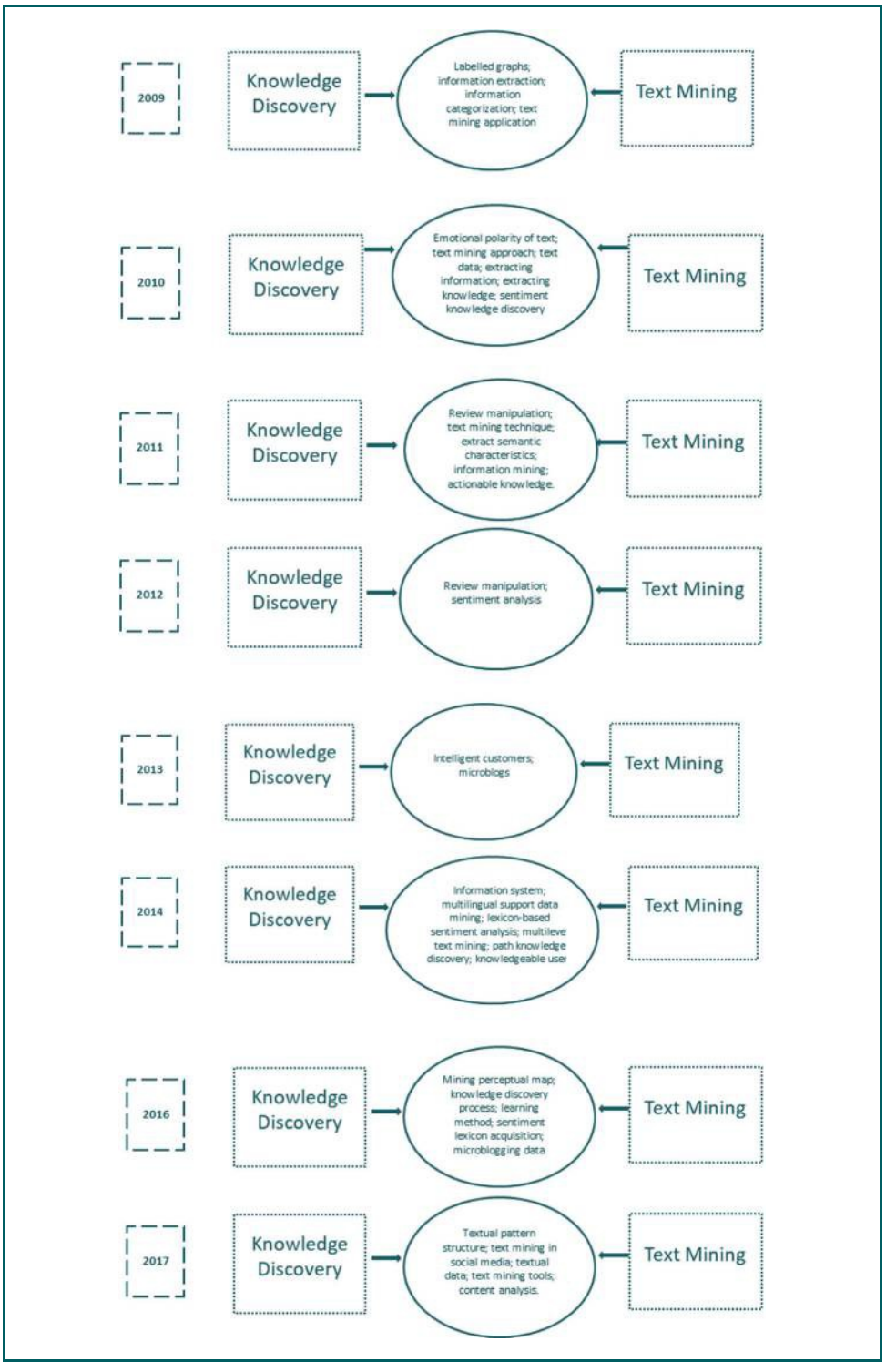

Figure 1 Second Screening Progress