Sinoara *et al. Journal of the Brazilian Computer Society* (2017) 23:9
DOI 10.1186/s13173-017-0058-7

Journal of the
Brazilian Computer Society

RESEARCH Open Access

Text mining and semantics: a systematic mapping study

Roberta Akemi Sinoara*, João Antunes and Solange Oliveira Rezende

Abstract

As text semantics has an important role in text meaning, the term semantics has been seen in a vast sort of text mining studies. However, there is a lack of studies that integrate the different research branches and summarize the developed works. This paper reports a systematic mapping about semantics-concerned text mining studies. This systematic mapping study followed a well-defined protocol. Its results were based on 1693 studies, selected among 3984 studies identified in five digital libraries. The produced mapping gives a general summary of the subject, points some areas that lacks the development of primary or secondary studies, and can be a guide for researchers working with semantics-concerned text mining. It demonstrates that, although several studies have been developed, the processing of semantic aspects in text mining remains an open research problem.

Keywords: Systematic review, Text mining, Text semantics

Introduction

Text mining techniques have become essential for supporting knowledge discovery as the volume and variety of digital text documents have increased, either in social networks and the Web or inside organizations. Text sources, as well as text mining applications, are varied. Although there is not a consensual definition established among the different research communities [1], text mining can be seen as a set of methods used to analyze unstructured data and discover patterns that were unknown beforehand [2].

A general text mining process can be seen as a five-step process, as illustrated in Fig. 1. The process starts with the specification of its objectives in the problem identification step. The text mining analyst, preferably working along with a domain expert, must delimit the text mining application scope, including the text collection that will be mined and how the result will be used. The specifications stated in the problem identification step will guide the next steps of the text mining process, which can be executed in cycles of data preparation (pre-processing step), knowledge discovery (pattern extraction step), and knowledge evaluation (post-processing step).

The pre-processing step is about preparing data for pattern extraction. In this step, raw text is transformed into some data representation format that can be used as input for the knowledge extraction algorithms. The activities performed in the pre-processing step are crucial for the success of the whole text mining process. The data representation must preserve the patterns hidden in the documents in a way that they can be discovered in the next step. In the pattern extraction step, the analyst applies a suitable algorithm to extract the hidden patterns. The algorithm is chosen based on the data available and the type of pattern that is expected. The extracted knowledge is evaluated in the post-processing step. If this knowledge meets the process objectives, it can be put available to the users, starting the final step of the process, the knowledge usage. Otherwise, another cycle must be performed, making changes in the data preparation activities and/or in pattern extraction parameters. If any changes in the stated objectives or selected text collection must be made, the text mining process should be restarted at the problem identification step.

Text data are not naturally in a format that is suitable for the pattern extraction, which brings additional challenges to an automatic knowledge discovery process. The meaning of natural language texts basically depends on

*Correspondence: rsinoara@usp.br
Laboratório de Inteligência Computacional (LABIC), Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP), P.O. Box 668, 13561-970 São Carlos, SP, Brazil

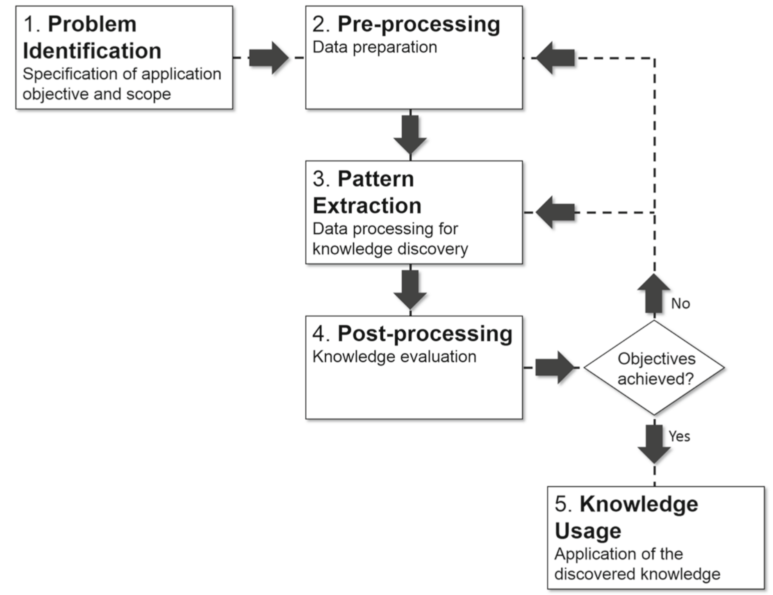

1. Problem Identification
Specification of application objective and scope
2. Pre-processing
Data preparation
3. Pattern Extraction
Data processing for knowledge discovery
4. Post-processing
Knowledge evaluation
Objectives achieved?
No
Yes
5. Knowledge Usage
Application of the discovered knowledge

Fig. 1 A general text mining process

lexical, syntactic, and semantic levels of linguistic knowledge. Each level is more complex and requires a more sophisticated processing than the previous level. This is a common trade-off when dealing with natural language processing: expressiveness versus processing cost. Thus, lexical and syntactic components have been more broadly explored in text mining than the semantic component [2]. Recently, text mining researchers have become more interested in text semantics, looking for improvements in the text mining results. The reason for this increasing interest can be assigned both to the progress of the computing capacity, which is constantly reducing the processing time, and to developments in the natural language processing field, which allow a deeper processing of raw texts.

In order to compare the expressiveness of each level of text interpretation (lexical, syntactic, and semantic), consider two simple sentences:

1. Company A acquired Company B.
2. Company B acquired Company A.

Sentences 1 and 2 have opposite meanings, but they have the same terms ("Company", "A", "B", "acquired"). Thus, if we analyze these sentences only in the lexical level, it is not possible to differentiate them. However, considering the sentence syntax, we can see that they are opposite. They have the same verb, and the subject of one sentence is the object of the other sentence and vice versa. If we analyze a little deeper, now considering the sentence semantics, we find that in sentence 1, "Company A" has the semantic role of *agent* regarding the verb "acquire" and "Company B" has the semantic role of *theme*. The same can be said to a third sentence:

3. Company B was acquired by Company A.

Despite the fact that syntactically sentences 1 and 3 have opposite subjects and objects, they have the same semantic roles. Thus, at the semantic level, they have the same meaning. If we go deeper and consider semantic relations among words (as the synonymy, for example), we can find that sentence 4 also expresses the same event:

4. Company A purchased Company B.

Besides, going even deeper in the interpretation of the sentences, we can understand their meaning—they are related to some takeover—and we can, for example, infer that there will be some impacts on the business environment.

Traditionally, text mining techniques are based on both a bag-of-words representation and application of data mining techniques. In this approach, only the lexical component of the texts are considered. In order to get a more complete analysis of text collections and get better text mining results, several researchers directed their attention to text semantics.

Text semantics can be considered in the three main steps of text mining process: pre-processing, pattern extraction and post-processing. In the pre-processing step, data representation can be based on some sort of semantic aspect of the text collection. In the pattern extraction, semantic information can be used to guide the model generation or to refine it. In the post-processing step, the extracted patterns can be evaluated based on semantic aspects. Either way, text mining based on text semantics can go further than text mining based only on lexicon or syntax. A proper treatment of text semantics can lead to more appropriate results for certain applications [2]. For example, semantic information has an important impact on document content and can be crucial to differentiate documents which, despite the use of the same vocabulary, present different ideas about the same subject.

The term semantics has been seen in a vast sort of text mining studies. However, there is a lack of studies that integrate the different branches of research performed to incorporate text semantics in the text mining process. Secondary studies, such as surveys and reviews, can integrate and organize the studies that were already developed and guide future works.

Thus, this paper reports a systematic mapping study to overview the development of semantics-concerned studies and fill a literature review gap in this broad research field through a well-defined review process. Semantics can be related to a vast number of subjects, and most of them are studied in the natural language processing field. As examples of semantics-related subjects, we can mention representation of meaning, semantic parsing and interpretation, word sense disambiguation, and coreference resolution. Nevertheless, the focus of this paper is not on semantics but on semantics-concerned text mining studies. As the term semantics appears in text mining studies in different contexts, this systematic mapping aims to present a general overview and point some areas that lack the development of primary studies and those areas that secondary studies would be of great help. This paper aims to point some directions to the reader who is interested in semantics-concerned text mining researches.

As it covers a wide research field, this systematic mapping study started with a space of 3984 studies, identified in five digital libraries. Due to time and resource limitations, except for survey papers, the mapping was done primarily through information found in paper abstracts. Therefore, our intention is to present an overview of semantics-concerned text mining, presenting a map of studies that has been developed by the research community, and not to present deep details of the studies. The papers were analyzed in relation to their application domains, performed tasks, applied methods and resources, and level of user's interaction. The contribution of this paper is threefold: (i) it presents an overview of semantics-concerned text mining studies from a text mining viewpoint, organizing the studies according to seven aspects (application domains, languages, external knowledge sources, tasks, methods and algorithms, representation models, and user's interaction); (ii) it quantifies and confirms some previous feelings that we had about our study subject; and (iii) it provides a starting point for those, researchers or practitioners, who are initiating works on semantics-concerned text mining.

The remainder of this paper is organized as follows. The "Method applied for systematic mapping" section presents an overview of systematic mapping method, since this is the type of literature review selected to develop this study and it is not widespread in the text mining community. In this section, we also present the protocol applied to conduct the systematic mapping study, including the research questions that guided this study and how it was conducted. The results of the systematic mapping, as well as identified future trends, are presented in the "Results and discussion" section. The "Conclusion" section concludes this work.

Method applied for systematic mapping

The review reported in this paper is the result of a systematic mapping study, which is a particular type of systematic literature review [3, 4]. Systematic literature review is a formal literature review adopted to identify, evaluate, and synthesize evidences of empirical results in order to answer a research question. It is extensively applied in medicine, as part of the evidence-based medicine [5]. This type of literature review is not as disseminated in the computer science field as it is in the medicine and health care fields[1], although computer science researches can also take advantage of this type of review. We can find important reports on the use of systematic reviews specially in the software engineering community [3, 4, 6, 7]. Other sparse initiatives can also be found in other computer science areas, as cloud-based environments [8], image pattern recognition [9], biometric authentication [10], recommender systems [11], and opinion mining [12].

A systematic review is performed in order to answer a research question and must follow a defined protocol. The protocol is developed when planning the systematic review, and it is mainly composed by the research questions, the strategies and criteria for searching for primary studies, study selection, and data extraction. The protocol is a documentation of the review process and must have all the information needed to perform the literature review in a systematic way. The analysis of selected studies, which is performed in the data extraction phase, will provide the answers to the research questions that motivated the literature review. Kitchenham and Charters [3] present a very useful guideline for planning and conducting systematic literature reviews. As systematic reviews follow a formal, well-defined, and documented protocol, they tend to be less biased and more reproducible than a regular literature review.

When the field of interest is broad and the objective is to have an overview of what is being developed in the research field, it is recommended to apply a particular type of systematic review named systematic mapping study [3, 4]. Systematic mapping studies follow an well-defined protocol as in any systematic review. The main differences between a traditional systematic review and a systematic mapping are their breadth and depth. While a systematic review deeply analyzes a low number of primary studies, in a systematic mapping a wider number of studies are analyzed, but less detailed. Thus, the search terms of a systematic mapping are broader and the results are usually presented through graphs. Systematic mapping studies can be used to get a mapping of the publications about some subject or field and identify areas that require the development of more primary studies and areas in which a narrower systematic literature review would be of great help to the research community.

This paper reports a systematic mapping study conducted to get a general overview of how text semantics is being treated in text mining studies. It fills a literature review gap in this broad research field through a well-defined review process. As a systematic mapping, our study follows the principles of a systematic mapping/review. However, as our goal was to develop a general mapping of a broad field, our study differs from the procedure suggested by Kitchenham and Charters [3] in two ways. Firstly, Kitchenham and Charters [3] state that the systematic review should be performed by two or more researchers. Although our mapping study was planned by two researchers, the study selection and the information extraction phases were conducted by only one due to the resource constraints. In this process, the other researchers reviewed the execution of each systematic mapping phase and their results. Secondly, systematic reviews usually are done based on primary studies only, nevertheless we have also accepted secondary studies (reviews or surveys) as we want an overview of all publications related to the theme.

In the following subsections, we describe our systematic mapping protocol and how this study was conducted.

Systematic mapping planning

The first step of a systematic review or systematic mapping study is its planning. The researchers conducting the study must define its protocol, i.e., its research questions and the strategies for identification, selection of studies, and information extraction, as well as how the study results will be reported. The main parts of the protocol that guided the systematic mapping study reported in this paper are presented in the following.

Research question: the main research question that guided this study was "How is semantics considered in text mining studies?" The main question was detailed in seven secondary questions, all of them related to text mining studies that consider text semantics in some way:

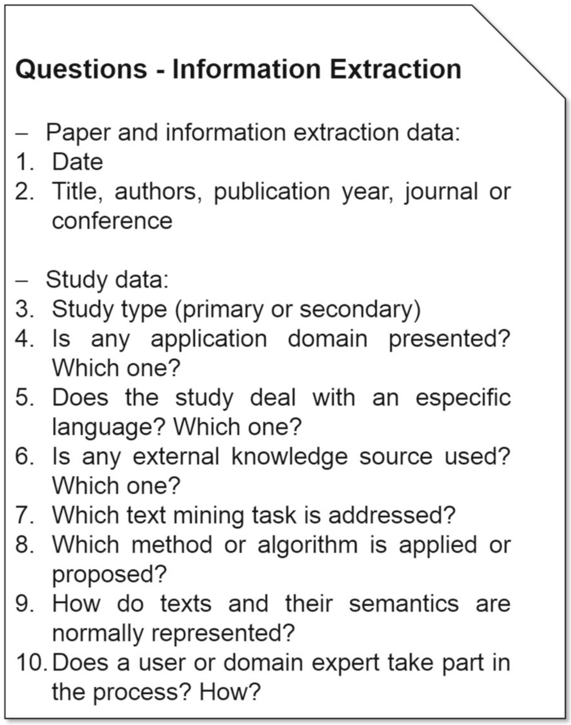
1. What are the application domains that focus on text semantics?
2. What are the natural languages being considered when working with text semantics?
3. Which external sources are frequently used in text mining studies when text semantics is considered?
4. In what text mining tasks is the text semantics most considered?
5. What methods and algorithms are commonly used?
6. How can texts be represented?
7. Do users or domain experts take part in the text mining process?

Study identification: the study identification was performed through searches for studies conducted in five digital libraries: ACM Digital Library, IEEE Xplore, Science Direct, Web of Science, and Scopus. The following general search expression was applied in both Title and Keywords fields, when allowed by the digital library search engine: `semantic* AND text* AND (mining OR representation OR clustering OR classification OR association rules)`.

Study selection: every study returned in the search phase went to the selection phase. Studies were selected based on title, abstract, and paper information (as number of pages, for example). Through this analysis, duplicated studies (most of them were studies found in more than one database) were identified. Besides, studies which match at least one of the following exclusion criteria were rejected: (i) one page papers, posters, presentations, abstracts, and editorials; (ii) papers hosted in services with restricted access and not accessible; (iii) papers written in languages different from English or Portuguese; and (iv) studies that do not deal with text mining and text semantics.

Information extraction: the information extraction phase was performed with papers accepted in the selection phase (papers that were not identified as duplicated or rejected). The abstracts were read in order to extract the information presented in Fig. 2.

As any literature review, this study has some bias. The advantage of a systematic literature review is that the protocol clearly specifies its bias, since the review process is well-defined. There are bias related to (i) study identification, i.e., only papers matching the search expression and returned by the searched digital libraries were selected; (ii) selection criteria, i.e., papers that matches the exclusion criteria were rejected; and (iii) information extraction, i.e., the information were mainly extracted considering only



Questions - Information Extraction

– Paper and information extraction data:
1. Date
2. Title, authors, publication year, journal or conference

– Study data:
3. Study type (primary or secondary)
4. Is any application domain presented? Which one?
5. Does the study deal with an especific language? Which one?
6. Is any external knowledge source used? Which one?
7. Which text mining task is addressed?
8. Which method or algorithm is applied or proposed?
9. How do texts and their semantics are normally represented?
10. Does a user or domain expert take part in the process? How?

Fig. 2 Information extraction form

title and abstracts. It is not feasible to conduct a literature review free of bias. However, it is possible to conduct it in a controlled and well-defined way through a systematic process.

Systematic mapping conduction

The conduction of this systematic mapping followed the protocol presented in the last subsection and is illustrated in Fig. 3. The selection and the information extraction phases were performed with support of the Start tool [13].

This paper reports the results obtained after the execution of two cycles of the systematic mapping phases. The first cycle was executed based on searches performed in January 2014. The second cycle was an update of the first cycle, with searches performed in February 2016[2]. A total of 3984 papers were found using the search expression in the five digital libraries. In the selection phase, 725 duplicated studies were identified and 1566 papers were rejected according to the exclusion criteria, mainly based on their title and abstract. Most of the rejected papers match the last exclusion criteria (*Studies that do not deal with text mining and text semantics*). Among them, we can find studies that deal with multimedia data (images, videos, or audio) and with construction, description, or annotation of corpus.

After the selection phase, 1693 studies were accepted for the information extraction phase. In this phase, information about each study was extracted mainly based on the abstracts, although some information was extracted from the full text. The results of the accepted paper mapping are presented in the next section.

Results and discussion

The mapping reported in this paper was conducted with the general goal of providing an overview of the researches developed by the text mining community and that are concerned about text semantics. This mapping is based on 1693 studies selected as described in the previous section. The distribution of these studies by publication year is presented in Fig. 4. We can note that text semantics has been addressed more frequently in the last years, when a higher number of text mining studies showed some interest in text semantics. The peak was in 2011, with 223 identified studies. The lower number of studies in the year 2016 can be assigned to the fact that the last searches were conducted in February 2016.

The results of the systematic mapping study is presented in the following subsections. We start our report presenting, in the "Surveys" section, a discussion about the eighteen secondary studies (surveys and reviews) that were identified in the systematic mapping. Then, each following section from "Application domains" to "User's interaction" is related to a secondary research question that guided our study, i.e., application domains, languages, external knowledge sources, text mining tasks, methods and algorithms, representation model, and user's interaction. In the "Systematic mapping summary and future trends" section, we present a consolidation of our results and point some gaps of both primary and secondary studies.

Some studies accepted in this systematic mapping are cited along the presentation of our mapping. We do not present the reference of every accepted paper in order to present a clear reporting of the results.

Surveys

In this systematic mapping, we identified 18 survey papers associated to the theme text mining and semantics [14–31]. Each paper exploits some particularity of this broad theme. In the following, we present a short overview of these papers, which is based on the full text of the papers.

Grobelnik [14] presents, briefly but in a very clear form, an interesting discussion of text processing in his three-page paper. The author organizes the field in three main dimensions, which can be used to classify text processing approaches: representation, technique, and task. The task dimension is about the kind of problems, we solve through the text processing. Document search, clustering, classification, summarization, trend detection, and monitoring are examples of tasks. Considering how text

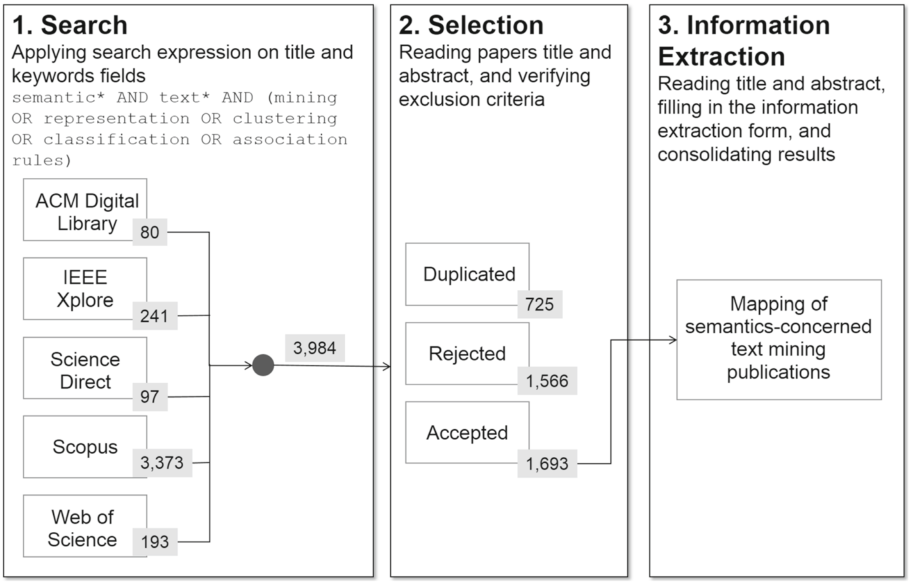

1. Search
Applying search expression on title and keywords fields
semantic* AND text* AND (mining OR representation OR clustering OR classification OR association rules)
ACM Digital Library
80
IEEE Xplore
241
Science Direct
97
Scopus
3,373
Web of Science
193
3,984
2. Selection
Reading papers title and abstract, and verifying exclusion criteria
Duplicated
725
Rejected
1,566
Accepted
1,693
3. Information Extraction
Reading title and abstract, filling in the information extraction form, and consolidating results
Mapping of semantics-concerned text mining publications

Fig. 3 Systematic mapping conduction phases. The numbers in the *shaded* areas indicate the quantity of studies involved

representations are manipulated (technique dimension), we have the methods and algorithms that can be used, including machine learning algorithms, statistical analysis, part-of-speech tagging, semantic annotation, and semantic disambiguation. In the representation dimension, we can find different options for text representation, such as words, phrases, bag-of-words, part-of-speech, subject-predicate-object triples and semantically annotated triples.

Grobelnik [14] also presents the levels of text representations, that differ from each other by the complexity of processing and expressiveness. The most simple level is the lexical level, which includes the common bag-of-words and n-grams representations. The next level is the syntactic level, that includes representations based on word co-location or part-of-speech tags. The most complete representation level is the semantic level and includes the representations based on word relationships,

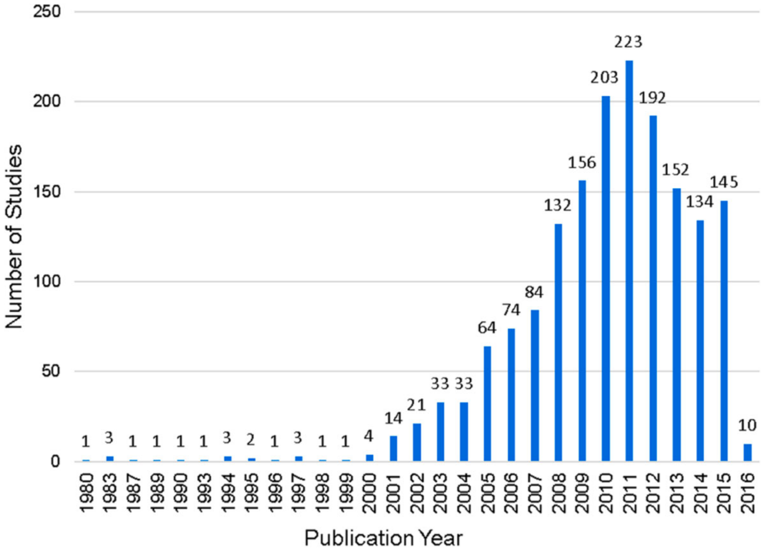


Fig. 4 Distribution of the 1693 accepted studies by publication year. Searches for studies identification were executed in January 2014 and February 2016

as the ontologies. Several different research fields deal with text, such as text mining, computational linguistics, machine learning, information retrieval, semantic web and crowdsourcing. Grobelnik [14] states the importance of an integration of these research areas in order to reach a complete solution to the problem of text understanding.

Stavrianou et al. [15] present a survey of semantic issues of text mining, which are originated from natural language particularities. This is a good survey focused on a linguistic point of view, rather than focusing only on statistics. The authors discuss a series of questions concerning natural language issues that should be considered when applying the text mining process. Most of the questions are related to text pre-processing and the authors present the impacts of performing or not some pre-processing activities, such as stopwords removal, stemming, word sense disambiguation, and tagging. The authors also discuss some existing text representation approaches in terms of features, representation model, and application task. The set of different approaches to measure the similarity between documents is also presented, categorizing the similarity measures by type (statistical or semantic) and by unit (words, phrases, vectors, or hierarchies).

Stavrianou et al. [15] also present the relation between ontologies and text mining. Ontologies can be used as background knowledge in a text mining process, and the text mining techniques can be used to generate and update ontologies. The authors conclude the survey stating that text mining is an open research area and that the objectives of the text mining process must be clarified before starting the data analysis, since the approaches must be chosen according to the requirements of the task being performed.

Methods that deal with latent semantics are reviewed in the study of Daud et al. [16]. The authors present a chronological analysis from 1999 to 2009 of directed probabilistic topic models, such as probabilistic latent semantic analysis, latent Dirichlet allocation, and their extensions. The models are classified according to their main functionality. They describe their advantages, disadvantages, and applications.

Wimalasuriya and Dou [17], Bharathi and Venkatesan [18], and Reshadat and Feizi-Derakhshi [19] consider the use of external knowledge sources (e.g., ontology or thesaurus) in the text mining process, each one dealing with a specific task. Wimalasuriya and Dou [17] present a detailed literature review of ontology-based information extraction. The authors define the recent information extraction subfield, named ontology-based information extraction (OBIE), identifying key characteristics of the OBIE systems that differentiate them from general information extraction systems. Besides, they identify a common architecture of the OBIE systems and classify existing systems along with different dimensions, as information extraction method applied, whether it constructs and updates the ontology, components of the ontology, and type of documents the system deals with. Bharathi and Venkatesan [18] present a brief description of several studies that use external knowledge sources as background knowledge for document clustering. Reshadat and Feizi-Derakhshi [19] present several semantic similarity measures based on external knowledge sources (specially WordNet and MeSH) and a review of comparison results from previous studies.

Schiessl and Bräscher [20] and Cimiano et al. [21] review the automatic construction of ontologies. Schiessl and Bräscher [20], the only identified review written in Portuguese, formally define the term ontology and discuss the automatic building of ontologies from texts. The authors state that automatic ontology building from texts is the way to the timely production of ontologies for current applications and that many questions are still open in this field. Also, in the theme of automatic building of ontologies from texts, Cimiano et al. [21] argue that automatically learned ontologies might not meet the demands of many possible applications, although they can already benefit several text mining tasks. The authors divide the ontology learning problem into seven tasks and discuss their developments. They state that ontology population task seems to be easier than learning ontology schema tasks.

Jovanovic et al. [22] discuss the task of semantic tagging in their paper directed at IT practitioners. Semantic tagging can be seen as an expansion of named entity recognition task, in which the entities are identified, disambiguated, and linked to a real-world entity, normally using a ontology or knowledge base. The authors compare 12 semantic tagging tools and present some characteristics that should be considered when choosing such type of tools.

Specifically for the task of irony detection, Wallace [23] presents both philosophical formalisms and machine learning approaches. The author argues that a model of the speaker is necessary to improve current machine learning methods and enable their application in a general problem, independently of domain. He discusses the gaps of current methods and proposes a pragmatic context model for irony detection.

The application of text mining methods in information extraction of biomedical literature is reviewed by Winnenburg et al. [24]. The paper describes the state-of-the-art text mining approaches for supporting manual text annotation, such as ontology learning, named entity and concept identification. They also describe and compare biomedical search engines, in the context of information retrieval, literature retrieval, result processing, knowledge retrieval, semantic processing, and integration of external tools. The authors argue that search engines must also be able to find results that are indirectly related to the user's keywords, considering the semantics and relationships between possible search results. They point that a good source for synonyms is WordNet.

Leser and Hakenberg [25] presents a survey of biomedical named entity recognition. The authors present the difficulties of both identifying entities (like genes, proteins, and diseases) and evaluating named entity recognition systems. They describe some annotated corpora and named entity recognition tools and state that the lack of corpora is an important bottleneck in the field.

Dagan et al. [26] introduce a special issue of the *Journal of Natural Language Engineering* on textual entailment recognition, which is a natural language task that aims to identify if a piece of text can be inferred from another. The authors present an overview of relevant aspects in textual entailment, discussing four PASCAL Recognising Textual Entailment (RTE) Challenges. They declared that the systems submitted to those challenges use cross-pair similarity measures, machine learning, and logical inference. The authors also describe tools, resources, and approaches commonly used in textual entailment tasks and conclude with the perspective that in the future, the constructed entailment "engines" will be used as a basic module by the text-understanding applications.

Irfan et al. [27] present a survey on the application of text mining methods in social network data. They present an overview of pre-processing, classification and clustering techniques to discover patterns from social networking sites. They point out that the application of text mining techniques can reveal patterns related to people's interaction behaviors. The authors present two basic pre-processing activities: feature extraction and feature selection. The authors also review classification and clustering approaches. They present different machine learning algorithms and discuss the importance of ontology usage to introduce explicit concepts, descriptions, and the semantic relationships among concepts. Irfan et al. [27] identify the main challenges related to the manipulation of social network texts (such as large data, data with impurities, dynamic data, emotions interpretations, privacy, and data confidence) and to text mining infrastructure (such as usage of cloud computing and improvement of the usability of text mining methods).

In the context of semantic web, Sheth et al. [28] define three types of semantics: implicit semantics, formal semantics, and powerful (or soft) semantics. Implicit semantics are those implicitly present in data patterns and is not explicitly represented in any machine processable syntax. Machine learning methods exploit this type of semantics. Formal semantics are those represented in some well-formed syntactic form and are machine-processable. The powerful semantics are the sort of semantics that allow uncertainty (that is, the representation of degree of membership and degree of certainty) and, therefore, allowing abductive or inductive reasoning. The authors also correlates the types of semantics with some core capabilities required by a practical semantic web application. The authors conclude their review asserting the importance of focusing research efforts in representation mechanisms for powerful semantics in order to move towards the development of semantic applications.

The formal semantics defined by Sheth et al. [28] is commonly represented by description logics, a formalism for knowledge representation. The application of description logics in natural language processing is the theme of the brief review presented by Cheng et al. [29].

The broad field of computational linguistics is presented by Martinez and Martinez [30]. Considering areas of computational linguistics that can be interesting to statisticians, the authors describe three main aspects of computational linguistics: formal language, information retrieval, and machine learning. The authors present common models for knowledge representation, addressing their statistical characteristics and providing an overview of information retrieval and machine learning methods related to computational linguistics. They describe some of the major statistical contributions to the areas of machine learning and computational linguistics, from the point of view of classification and clustering algorithms. Martinez and Martinez [30] emphasize that machine translation, part-of-speech tagging, word sense disambiguation, and text summarization are some of the identified applications that statisticians can contribute.

Bos [31] presents an extensive survey of computational semantics, a research area focused on computationally understanding human language in written or spoken form. He discusses how to represent semantics in order to capture the meaning of human language, how to construct these representations from natural language expressions, and how to draw inferences from the semantic representations. The author also discusses the generation of background knowledge, which can support reasoning tasks. Bos [31] indicates machine learning, knowledge resources, and scaling inference as topics that can have a big impact on computational semantics in the future.

As presented in this section, the reviewed secondary studies exploit some specific issues of semantics-concerned text mining researches. In contrast to them, this paper reviews a broader range of text mining studies that deal with semantic aspects. To the best of our knowledge, this is the first report of a mapping of this field. We present the results of our systematic mapping study in the following sections, organized in seven dimensions of the text mining studies derived from our secondary research questions: application domains, languages, external knowledge usage, tasks, methods and algorithms, representation model, and user's interaction.

Application domains

Research question:

What are the application domains that focus on text semantics?

Figure 5 presents the domains where text semantics is most present in text mining applications. Health care and life sciences is the domain that stands out when talking about text semantics in text mining applications. This fact is not unexpected, since life sciences have a long time concern about standardization of vocabularies and taxonomies. The building of taxonomies and ontologies is such a common practice in health care and life sciences that World Wide Web Consortium (W3C) has an interest group specific for developing, evaluating, and supporting semantic web technologies for this field [32]. Among the most common problems treated through the use of text mining in the health care and life science is the information retrieval from publications of the field. The search engine PubMed [33] and the MEDLINE database are the main text sources among these studies. There are also studies related to the extraction of events, genes, proteins and their associations [34–36], detection of adverse drug reaction [37], and the extraction of cause-effect and disease-treatment relations [38–40].

The second most frequent identified application domain is the mining of web texts, comprising web pages, blogs, reviews, web forums, social medias, and email filtering [41–46]. The high interest in getting some knowledge from web texts can be justified by the large amount and diversity of text available and by the difficulty found in manual analysis. Nowadays, any person can create content in the web, either to share his/her opinion about some product or service or to report something that is taking place in his/her neighborhood. Companies, organizations, and researchers are aware of this fact, so they are increasingly interested in using this information in their favor. Some competitive advantages that business can gain from the analysis of social media texts are presented in [47–49]. The authors developed case studies demonstrating how text mining can be applied in social media intelligence. From our systematic mapping data, we found that Twitter is the most popular source of web texts and its posts are commonly used for sentiment analysis or event extraction.

Besides the top 2 application domains, other domains that show up in our mapping refers to the mining of specific types of texts. We found research studies in mining news, scientific papers corpora, patents, and texts with economic and financial content.

Languages

Research question:

What are the natural languages being considered when working with text semantics?

Whether using machine learning or statistical techniques, the text mining approaches are usually language independent. However, specially in the natural language processing field, annotated corpora is often required to train models in order to resolve a certain task for each specific language (semantic role labeling problem is an example). Besides, linguistic resources as semantic networks or lexical databases, which are language-specific, can be used to enrich textual data. Most of the resources available are English resources. Thus, the low number of annotated data or linguistic resources can be a bottleneck when working with another language. There are important initiatives to the development of researches for other languages, as an example, we have the ACM Transactions on Asian and Low-Resource Language Information Processing [50], an ACM journal specific for that subject.

In this study, we identified the languages that were mentioned in paper abstracts. The collected data are summarized in Fig. 6. We must note that English can be seen as

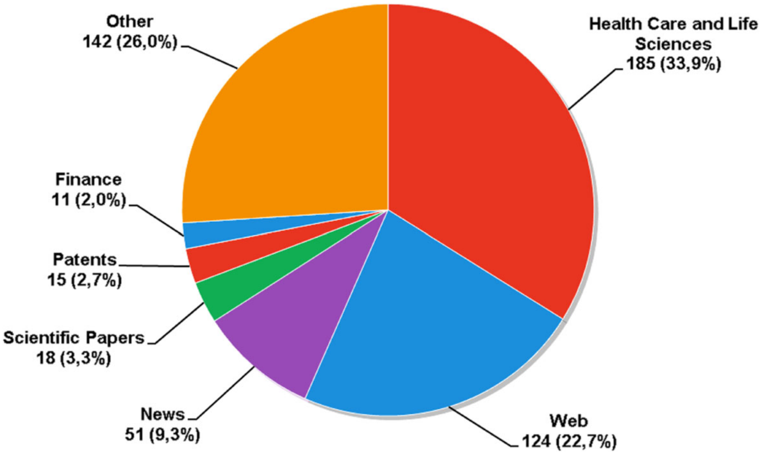


Fig. 5 Application domains identified in the literature mapping accepted studies

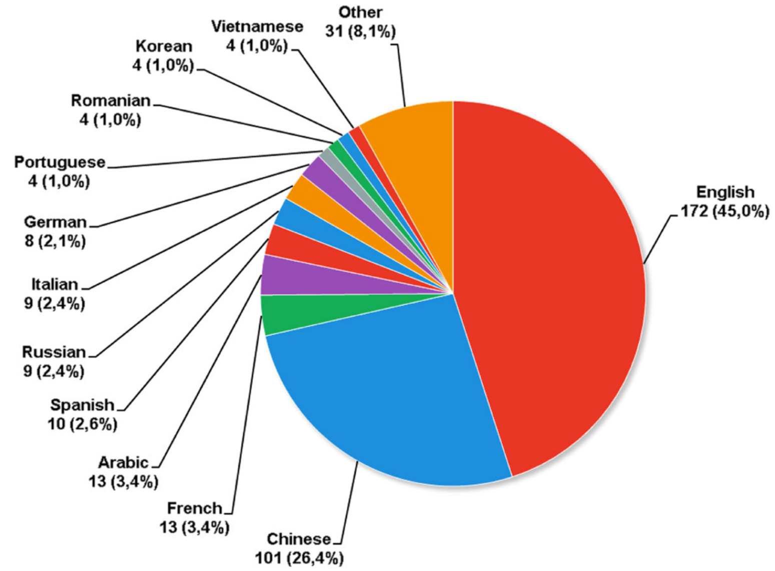


Fig. 6 Languages identified in the literature mapping accepted studies

a standard language in scientific publications; thus, papers whose results were tested only in English datasets may not mention the language, as examples, we can cite [51–56]. Besides, we can find some studies that do not use any linguistic resource and thus are language independent, as in [57–61]. These facts can justify that English was mentioned in only 45.0% of the considered studies.

Chinese is the second most mentioned language (26.4% of the studies reference the Chinese language). Wu et al. [62] point two differences between English and Chinese: in Chinese, there are no white spaces between words in a sentence and there are a higher number of frequent words (the number of frequent words in Chinese is more than twice the number of English frequent words). These characteristics motivate the development of methods and experimental evaluations specifically for Chinese.

This mapping shows that there is a lack of studies considering languages other than English or Chinese. The low number of studies considering other languages suggests that there is a need for construction or expansion of language-specific resources (as discussed in "External knowledge sources" section). These resources can be used for enrichment of texts and for the development of language specific methods, based on natural language processing.

External knowledge sources

Research question:

Which external sources are frequently used in text mining studies when text semantics is considered?

Text mining initiatives can get some advantage by using external sources of knowledge. Thesauruses, taxonomies, ontologies, and semantic networks are knowledge sources that are commonly used by the text mining community. Semantic networks is a network whose nodes are concepts that are linked by semantic relations. The most popular example is the WordNet [63], an electronic lexical database developed at the Princeton University. Depending on its usage, WordNet can also be seen as a thesaurus or a dictionary [64].

There is not a complete definition for the terms thesaurus, taxonomy, and ontology that is unanimously accepted by all research areas. Weller [65] presents an interesting discussion about the term *ontology*, including its origin and proposed definitions. She concluded the discussion stating that: "Ontologies should unambiguously represent shared background knowledge that helps people within a community of interest to understand each other. And they should make computer-readable indexing of information possible on the Web" [65]. The same can be said about thesauruses and taxonomies. In a general way, thesauruses, taxonomies, and ontologies are normally specialized in a specific domain and they usually differs from each other by their degree of expressiveness and complexity in their relational constructions [66]. Ontology would be the most expressive type of knowledge representation, having the most complex relations and formalized construction.

When looking at the external knowledge sources used in semantics-concerned text mining studies (Fig. 7), WordNet is the most used source. This lexical resource is cited by 29.9% of the studies that uses information beyond the text data. WordNet can be used to create or expand the current set of features for subsequent text classification

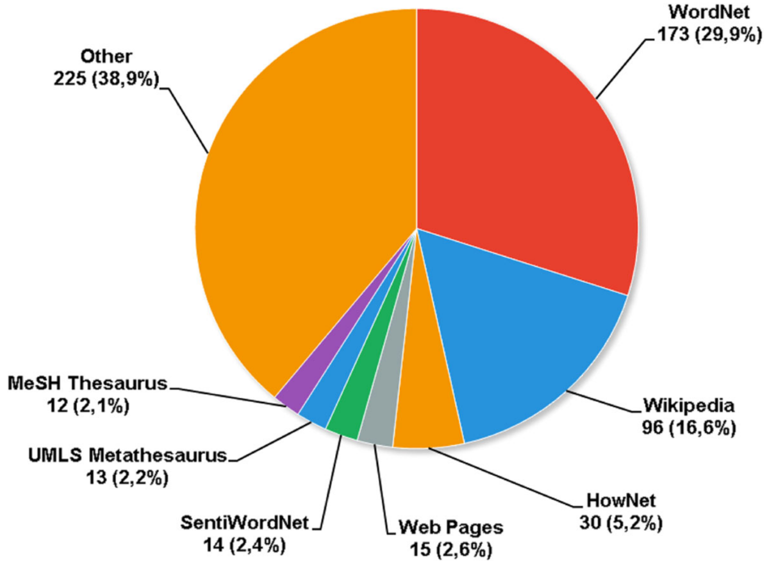


Fig. 7 External sources identified in the literature mapping accepted studies

or clustering. The use of features based on WordNet has been applied with and without good results [55, 67–69]. Besides, WordNet can support the computation of semantic similarity [70, 71] and the evaluation of the discovered knowledge [72].

The second most used source is Wikipedia [73], which covers a wide range of subjects and has the advantage of presenting the same concept in different languages. Wikipedia concepts, as well as their links and categories, are also useful for enriching text representation [74–77] or classifying documents [78–80]. Medelyan et al. [81] present the value of Wikipedia and discuss how the community of researchers are making use of it in natural language processing tasks (in special word sense disambiguation), information retrieval, information extraction, and ontology building.

The use of Wikipedia is followed by the use of the Chinese-English knowledge database HowNet [82]. Finding HowNet as one of the most used external knowledge source it is not surprising, since Chinese is one of the most cited languages in the studies selected in this mapping (see the "Languages" section). As well as WordNet, HowNet is usually used for feature expansion [83–85] and computing semantic similarity [86–88].

Web pages are also used as external sources [89–91]. Normally, web search results are used to measure similarity between terms. We also found some studies that use SentiWordNet [92], which is a lexical resource for sentiment analysis and opinion mining [93, 94]. Among other external sources, we can find knowledge sources related to Medicine, like the UMLS Metathesaurus [95–98], MeSH thesaurus [99–102], and the Gene Ontology [103–105].

Text mining tasks

Research question:

In what text mining tasks is the text semantics most considered?

The distribution of text mining tasks identified in this literature mapping is presented in Fig. 8. Classification and clustering are the most frequent tasks. Classification corresponds to the task of finding a model from examples with known classes (labeled instances) in order to predict the classes of new examples. On the other hand, clustering is the task of grouping examples (whose classes are unknown) based on their similarities. Classification was identified in 27.4% and clustering in 17.0% of the studies. As these are basic text mining tasks, they are often the basis of other more specific text mining tasks, such as sentiment analysis and automatic ontology building. Therefore, it was expected that classification and clustering would be the most frequently applied tasks.

Besides classification and clustering, we can note that semantic concern are present in tasks as information extraction [106–108], information retrieval [109–111], sentiment analysis [112–115], and automatic ontology building [116, 117], as well as the pre-processing step itself [118, 119].

Methods and algorithms

Research question:

What methods and algorithms are commonly used?

A word cloud[3] of methods and algorithms identified in this literature mapping is presented in Fig. 9, in which the font size reflects the frequency of the methods and algorithms among the accepted papers. We can note that

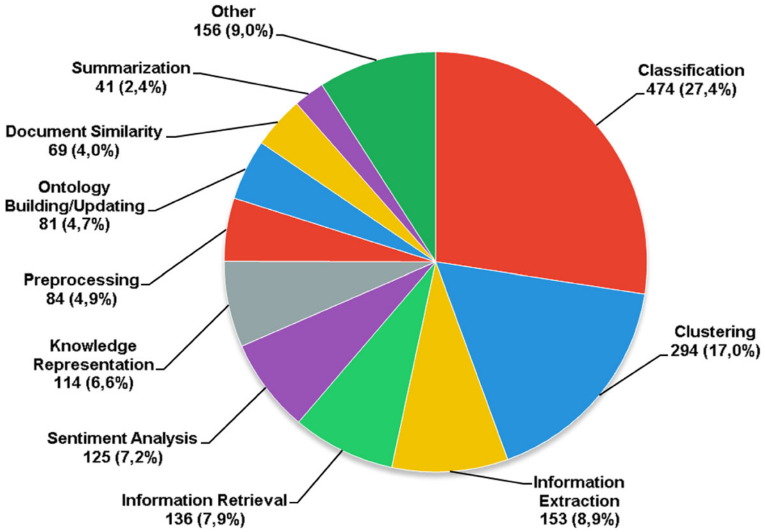


Fig. 8 Text mining tasks identified in the literature mapping accepted studies

the most common approach deals with latent semantics through Latent Semantic Indexing (LSI) [2, 120], a method that can be used for data dimension reduction and that is also known as latent semantic analysis. The Latent Semantic Index low-dimensional space is also called semantic space. In this semantic space, alternative forms expressing the same concept are projected to a common representation. It reduces the noise caused by synonymy and polysemy; thus, it latently deals with text semantics. Another technique in this direction that is commonly used for topic modeling is latent Dirichlet allocation (LDA) [121]. The topic model obtained by LDA has been used for representing text collections as in [58, 122, 123].

Beyond latent semantics, the use of concepts or topics found in the documents is also a common approach. The concept-based semantic exploitation is normally based on external knowledge sources (as discussed in the "External knowledge sources" section) [74, 124–128]. As an example, explicit semantic analysis [129] rely on Wikipedia to represent the documents by a concept vector. In a similar way, Spanakis et al. [125] improved hierarchical clustering quality by using a text representation based on concepts and other Wikipedia features, such as links and categories.

The issue of text ambiguity has also been the focus of studies. Word sense disambiguation can contribute to a

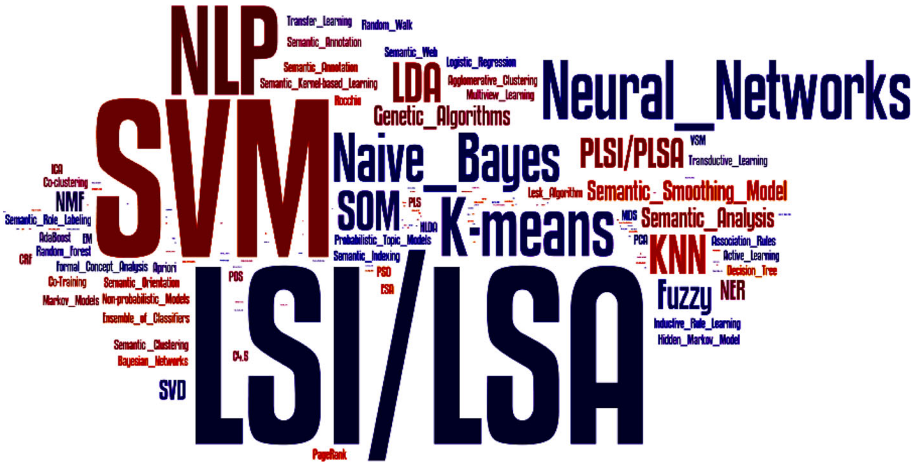


Fig. 9 Word cloud of methods and algorithms identified in the literature mapping studies. To enable a better reading of the word cloud, the frequency of the methods and algorithms higher than one was rounded up to the nearest ten (for example, a method applied in 75 studies is represented in the word cloud in a word size corresponding to the frequency 80)

better document representation. It is normally based on external knowledge sources and can also be based on machine learning methods [36, 130–133].

Other approaches include analysis of verbs in order to identify relations on textual data [134–138]. However, the proposed solutions are normally developed for a specific domain or are language dependent.

In Fig. 9, we can observe the predominance of traditional machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, K-means, and k-Nearest Neighbors (KNN), in addition to artificial neural networks and genetic algorithms. The application of natural language processing methods (NLP) is also frequent. Among these methods, we can find named entity recognition (NER) and semantic role labeling. It shows that there is a concern about developing richer text representations to be input for traditional machine learning algorithms, as we can see in the studies of [55, 139–142].

Text representation models

Research question:

How can texts be represented?

The most popular text representation model is the vector space model. In this model, each document is represented by a vector whose dimensions correspond to features found in the corpus. When features are single words, the text representation is called bag-of-words. Despite the good results achieved with a bag-of-words, this representation, based on independent words, cannot express word relationships, text syntax, or semantics. Therefore, it is not a proper representation for all possible text mining applications.

The use of richer text representations is the focus of several studies [62, 79, 97, 143–148]. Most of the studies concentrate on proposing more elaborated features to represent documents in the vector space model, including the use of topic model techniques, such as LSI and LDA, to obtain latent semantic features. Deep learning [149] is currently applied to represent independent terms through their associated concepts, in an attempt to narrow the relationships between the terms [150, 151]. The use of distributed word representations (word embeddings) can be seen in several works of this area in tasks such as classification [88, 152, 153], summarization [154], and information retrieval [155].

Besides the vector space model, there are text representations based on networks (or graphs), which can make use of some text semantic features. Network-based representations, such as bipartite networks and co-occurrence networks, can represent relationships between terms or between documents, which is not possible through the vector space model [147, 156–158].

In addition to the text representation model, text semantics can also be incorporated to text mining process through the use of external knowledge sources, like semantic networks and ontologies, as discussed in the "External knowledge sources" section.

User's interaction

Research question:

Do users or domain experts take part in the text mining process?

Text mining is a process to automatically discover knowledge from unstructured data. Nevertheless, it is also an interactive process, and there are some points where a user, normally a domain expert, can contribute to the process by providing his/her previous knowledge and interests. As an example, in the pre-processing step, the user can provide additional information to define a stoplist and support feature selection. In the pattern extraction step, user's participation can be required when applying a semi-supervised approach. In the post-processing step, the user can evaluate the results according to the expected knowledge usage.

Despite the fact that the user would have an important role in a real application of text mining methods, there is not much investment on user's interaction in text mining research studies. A probable reason is the difficulty inherent to an evaluation based on the user's needs. In empirical research, researchers use to execute several experiments in order to evaluate proposed methods and algorithms, which would require the involvement of several users, therefore making the evaluation not feasible in practical ways.

Less than 1% of the studies that were accepted in the first mapping cycle presented information about requiring some sort of user's interaction in their abstract. To better analyze this question, in the mapping update performed in 2016, the full text of the studies were also considered. Figure 10 presents types of user's participation

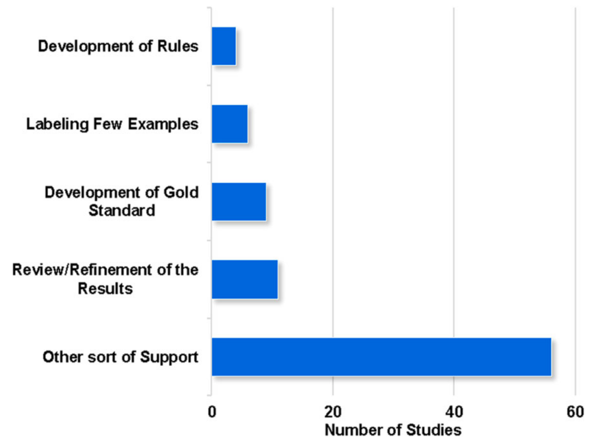


Fig. 10 Types of user participation identified in the literature mapping accepted studies

identified in the literature mapping studies. The most common user's interactions are the revision or refinement of text mining results [159–161] and the development of a standard reference, also called as gold standard or ground truth, which is used to evaluate text mining results [162–165]. Besides that, users are also requested to manually annotate or provide a few labeled data [166, 167] or generate of hand-crafted rules [168, 169].

Systematic mapping summary and future trends

Research question:

How is semantics considered in text mining studies?

Semantics is an important component in natural language texts. Consequently, in order to improve text mining results, many text mining researches claim that their solutions treat or consider text semantics in some way. However, text mining is a wide research field and there is a lack of secondary studies that summarize and integrate the different approaches. How is semantics considered in text mining studies? Looking for the answer to this question, we conducted this systematic mapping based on 1693 studies, accepted among the 3984 studies identified in five digital libraries. In the previous subsections, we presented the mapping regarding to each secondary research question. In this subsection, we present a consolidation of our results and point some future trends of semantics-concerned text mining.

As previously stated, the objective of this systematic mapping is to provide a general overview of semantics-concerned text mining studies. The papers considered in this systematic mapping study, as well as the mapping results, are limited by the applied search expression and the research questions. It is not feasible to cover all published papers in this broad field. Therefore, the reader can miss in this systematic mapping report some previously known studies. It is not our objective to present a detailed survey of every specific topic, method, or text mining task. This systematic mapping is a starting point, and surveys with a narrower focus should be conducted for reviewing the literature of specific subjects, according to one's interests.

The quantitative analysis of the scientific production by each text mining dimension (presented from the "Application domains" section to the "User's interaction" section) confirmed some previous feelings that we had about our study subject and highlighted other interesting characteristics of the field. Text semantics is closely related to ontologies and other similar types of knowledge representation. We also know that health care and life sciences is traditionally concerned about standardization of their concepts and concepts relationships. Thus, as we already expected, health care and life sciences was the most cited application domain among the literature accepted studies. This application domain is followed by the Web domain, what can be explained by the constant growth, in both quantity and coverage, of Web content.

It was surprising to find the high presence of the Chinese language among the studies. Chinese language is the second most cited language, and the HowNet, a Chinese-English knowledge database, is the third most applied external source in semantics-concerned text mining studies. Looking at the languages addressed in the studies, we found that there is a lack of studies specific to languages other than English or Chinese. We also found an expressive use of WordNet as an external knowledge source, followed by Wikipedia, HowNet, Web pages, SentiWordNet, and other knowledge sources related to Medicine.

Text classification and text clustering, as basic text mining tasks, are frequently applied in semantics-concerned text mining researches. Among other more specific tasks, sentiment analysis is a recent research field that is almost as applied as information retrieval and information extraction, which are more consolidated research areas. SentiWordNet, a lexical resource for sentiment analysis and opinion mining, is already among the most used external knowledge sources.

The treatment of latent semantics, through the application of LSI, stands out when looking at methods and algorithms. Besides that, traditional text mining methods and algorithms, like SVM, KNN, and K-means, are frequently applied and researches tend to enhance the text representation by applying NLP methods or using external knowledge sources. Thus, text semantics can be incorporated to the text mining process mainly through two approaches: the construction of richer terms in the vector space representation model or the use of networks or graphs to represent semantic relations between terms or documents.

In real application of the text mining process, the participation of domain experts can be crucial to its success. However, the participation of users (domain experts) is seldom explored in scientific papers. The difficulty inherent to the evaluation of a method based on user's interaction is a probable reason for the lack of studies considering this approach.

The mapping indicates that there is space for secondary studies in areas that has a high number of primary studies, such as studies of feature enrichment for a better text representation in the vector space model; use of classification methods; use of clustering methods; and the use of latent semantics in text mining. A detailed literature review, as the review of Wimalasuriya and Dou [17] (described in "Surveys" section), would be worthy for organization and summarization of these specific research subjects.

Considering the development of primary studies, we identified three main future trends: user's interaction, non-English text processing, and graph-based representation. We expect an increase in the number of studies that have some level of user's interaction to bring his/her needs and interests to the process. This is particularly valuable for the clustering task, because a considered good clustering solution can vary from user to user [170]. We also expect a raise of resources (linguistic resources and annotated corpora) for non-English languages. These resources are very important to the development of semantics-concerned text mining techniques. Higher availability of non-English resources will allow a higher number of studies dealing with these languages. Another future trend is the development and use of graph-based text representation. Nowadays, there are already important researches in this direction, and we expect that it will increase as graph-based representations are more expressive than traditional representations in the vector space model.

As an alternative summary of this systematic mapping, additional visualizations of both the selected studies and systematic mapping results can be found online at http://sites.labic.icmc.usp.br/pinda_sm. For this purpose, the prototype of the Pinda tool was adapted for hierarchical visualization of the textual data, using K-means algorithm to group the results. The tool allows the analysis of data (title + abstract of selected studies or information extracted from them) through multiple visualization techniques (Thumbnail, Snippets, Directories, Scatterplot, Treemap, and Sunburst), coordinating the user's interactions for a better understanding of existing relationships. Figure 11 illustrates the Scatterplot visualization of studies accepted in this systematic mapping. Some of the possible visualizations of the systematic mapping results are presented in Fig. 12.

Conclusion

Text semantics are frequently addressed in text mining studies, since it has an important influence in text meaning. However, there is a lack of secondary studies that consolidate these researches. This paper reported a systematic mapping study conducted to overview semantics-concerned text mining literature. The scope of this mapping is wide (3984 papers matched the search expression). Thus, due to limitations of time and resources, the mapping was mainly performed based on abstracts of papers. Nevertheless, we believe that our limitations do not have a crucial impact on the results, since our study has a broad coverage.

The main contributions of this work are (i) it presents a quantitative analysis of the research field; (ii) its conduction followed a well-defined literature review protocol; (iii) it discusses the area regarding seven important text mining dimensions: application domain, language, external knowledge source, text mining task, method and algorithm, representation model, and user's interaction; and (iv) the produced mapping can give a general summary of the subject and can be of great help for researchers working with semantics and text mining. Thus, this work filled a gap in the literature as, to the best of our knowledge, this is the first general literature review of this wide subject.

Although several researches have been developed in the text mining field, the processing of text semantics remains an open research problem. The field lacks secondary studies in areas that has a high number of primary studies, such as feature enrichment for a better text representation in the vector space model. Another highlight is about a language-related issue. We found considerable differences in numbers of studies among different languages, since 71.4% of the identified studies deal with English and Chinese. Thus, there is a lack of studies dealing with texts written in other languages. When considering semantics-concerned text mining, we believe that this lack can be filled with the development of good knowledge bases and natural language processing methods specific for these languages. Besides, the analysis of the impact of languages in semantic-concerned text mining is also an interesting open research question. A comparison among semantic aspects of different languages and their impact on the results of text mining techniques would also be interesting.

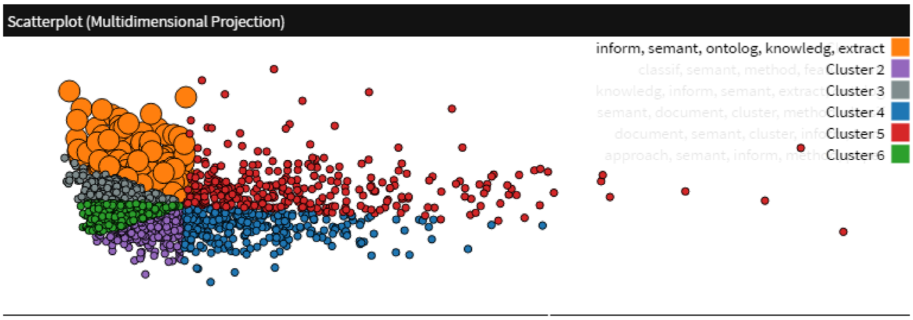

Scatterplot (Multidimensional Projection)
inform, semant, ontolog, knowledg, extract
classif, semant, method, fea Cluster 2
knowledg, inform, semant, extract Cluster 3
semant, document, cluster, metho Cluster 4
document, semant, cluster, info Cluster 5
approach, semant, inform, metho Cluster 6

Fig. 11 Scatterplot visualization of accepted studies of the systematic mapping

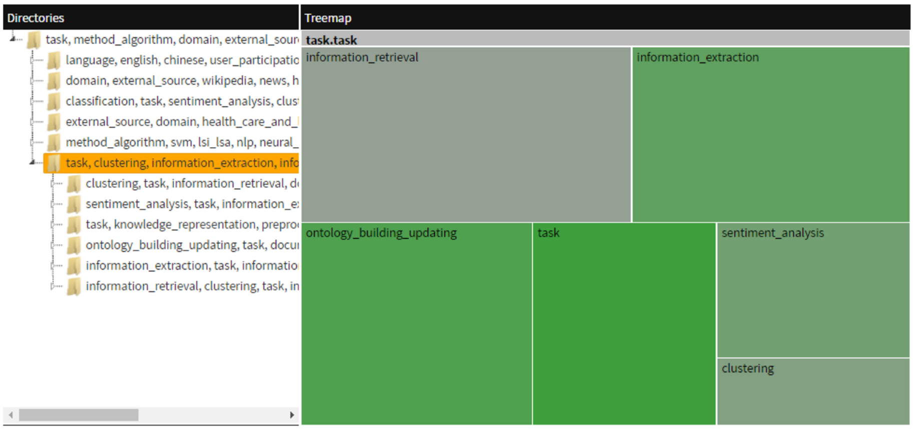

Directories
Treemap
task, method_algorithm, domain, external_sour
language, english, chinese, user_participatic
domain, external_source, wikipedia, news, h
classification, task, sentiment_analysis, clus
external_source, domain, health_care_and_
method_algorithm, svm, lsi_lsa, nlp, neural_
task, clustering, information_extraction, info
clustering, task, information_retrieval, d
sentiment_analysis, task, information_e
task, knowledge_representation, prepro
ontology_building_updating, task, docu
information_extraction, task, informatio
information_retrieval, clustering, task, in
task.task
information_retrieval
information_extraction
ontology_building_updating
task
sentiment_analysis
clustering

Fig. 12 Directories and Treemap visualizations of the systematic mapping results

Endnotes

[1] A simple search for "systematic review" on the Scopus database in June 2016 returned, by subject area, 130,546 Health Sciences documents (125,254 of them for Medicine) and only 5,539 Physical Sciences (1328 of them for Computer Science). The coverage of Scopus publications are balanced between Health Sciences (32% of total Scopus publication) and Physical Sciences (29% of total Scopus publication).

[2] It was not possible to perform the second cycle of searches in ACM Digital Library because of a change in the interface of this search engine. However, it must be notice that only eight studies that was found only in this database was accepted in the first cycle. All other studies was also retrieved by other search engines (specially Scopus, which retrieved more than 89% of accepted studies.)

[3] Word cloud created with support of Wordle [171].

Acknowledgements

The authors would like to thank the financial support of grant #132666/2016-2, National Council for Scientific and Technological Development (CNPq); grants #2013/14757-6, #2014/08996-0, and #2016/07620-2, São Paulo Research Foundation (FAPESP); and Coordination for the Improvement of Higher Education Personnel (CAPES).

Authors' contributions

RAS and SOR planned this systematic mapping study. RAS conducted its first cycle (searches performed in January 2014). JA and RAS conducted its second cycle (searches performed in February 2016). RAS and SOR analyzed the results and drafted the manuscript after the first cycle and updated it after the second cycle. JA was involved in updating the manuscript with the second cycle results. All authors revised and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 24 March 2017 Accepted: 1 June 2017
Published online: 29 June 2017

References

1. Miner G, Elder J, Hill T, Nisbet R, Delen D, Fast A (2012) Practical text mining and statistical analysis for non-structured text data applications. 1st edn. Academic Press, Boston
2. Aggarwal CC, Zhai C (eds) (2012) Mining text data. Springer, Durham
3. Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01. Keele University and Durham University Joint Report, Durham, UK
4. Petersen K, Feldt R, Mujtaba S, Mattsson M (2008) Systematic mapping studies in software engineering. In: EASE 2008: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering. EASE'08. British Computer Society, Swinton, UK. pp 68–77
5. Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M (2007) Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 80(4):571–583
6. Kitchenham B, Pretorius R, Budgen D, Brereton OP, Turner M, Niazi M, et al (2010) Systematic literature reviews in software engineering—a tertiary study. Inf Softw Technol 52(8):792–805
7. Felizardo KR, Nakagawa EY, MacDonell SG, Maldonado JC (2014) A visual analysis approach to update systematic reviews. In: EASE'14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. ACM, New York. pp 4:1–4:10
8. Moghaddam FA, Lago P, Grosso P (2015) Energy-efficient networking solutions in cloud-based environments: a systematic literature review. ACM Comput Surv 47(4):64:1–64:32
9. Pedro RWD, Nunes FLS, Machado-Lima A (2013) Using grammars for pattern recognition in images: a systematic review. ACM Comput Surv 46(2):26:1–26:34
10. Pisani PH, Lorena AC (2013) A systematic review on keystroke dynamics. J Braz Comput Soc 19(4):573–587
11. Park DH, Kim HK, Choi IY, Kim JK (2012) A literature review and classification of recommender systems research. Expert Syst Appl 39(11):10059–10072
12. Khan K, Baharudin BB, Khan A, et al (2009) Mining opinion from text documents: a survey. In: DEST'09: Proceedings of the 3rd IEEE International Conference on Digital Ecosystems and Technologies. IEEE. pp 217–222
13. Laboratory of Research on Software Engineering (LaPES) - StArt Tool. http://lapes.dc.ufscar.br/tools/start_tool. Accessed 8 June 2016
14. Grobelnik M (2011) Many faces of text processing. In: WIMS'11: Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM. p 5
15. Stavrianou A, Andritsos P, Nicoloyannis N (2007) Overview and semantic issues of text mining. SIGMOD Rec 36(3):23–34

16. Daud A, Li J, Zhou L, Muhammad F (2010) Knowledge discovery through directed probabilistic topic models: a survey. Front Comput Sci China 4(2):280–301
17. Wimalasuriya DC, Dou D (2010) Ontology-based information extraction: an introduction and a survey of current approaches. J Inf Sci 36(3):306–323
18. Bharathi G, Venkatesan D (2012) Study of ontology or thesaurus based document clustering and information retrieval. J Eng Appl Sci 7(4):342–347
19. Reshadat V, Feizi-Derakhshi MR (2012) Studying of semantic similarity methods in ontology. Res J Appl Sci Eng Technol 4(12):1815–1821
20. Schiessl M, Bräscher M (2012) Do texto às ontologias: uma perspectiva para a ciência da informação. Ciência da Informação 40(2):301–311
21. Cimiano P, Völker J, Studer R (2006) Ontologies on demand?—a description of the state-of-the-art, applications, challenges and trends for ontology learning from text. Inf Wiss Prax 57(6-7):315–320
22. Jovanovic J, Bagheri E, Cuzzola J, Gasevic D, Jeremic Z, Bashash R (2014) Automated semantic tagging of textual content. IT Prof 16(6):38–46
23. Wallace BC (2015) Computational irony: a survey and new perspectives. Artif Intell Rev 43(4):467–483
24. Winnenburg R, Wächter T, Plake C, Doms A, Schroeder M (2008) Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? Brief Bioinform 9(6):466–478
25. Leser U, Hakenberg J (2005) What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform 6(4):357–369
26. Dagan I, Dolan B, Magnini B, Roth D (2009) Recognizing textual entailment: rational, evaluation and approaches. Nat Lang Eng 15(04):i–xvii
27. Irfan R, King CK, Grages D, Ewen S, Khan SU, Madani SA, et al. (2015) A survey on text mining in social networks. Knowl Eng Rev 30(02):157–170
28. Sheth A, Ramakrishnan C, Thomas C (2005) Semantics for the semantic web: the implicit, the formal and the powerful. Int J Semant Web Inf Syst 1(1):1–18
29. Cheng XY, Cheng C, Zhu Q (2011) The applications of description logics in natural language processing. Adv Mater Res 204:381–386
30. Martinez A, Martinez W (2015) At the interface of computational linguistics and statistics. Wiley Interdiscip Rev Comput Stat 7(4):258–274
31. Bos J (2011) A survey of computational semantics: representation, inference and knowledge in wide-coverage text understanding. Lang Linguist Compass 5(6):336–366
32. W3C - Semantic Web Health Care and Life Sciences Interest Group. https://www.w3.org/blog/hcls/. Accessed 8 June 2016
33. National Center for Biotechnology Information - PubMed. http://www.ncbi.nlm.nih.gov/pubmed/. Accessed 8 June 2016
34. Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S (2012) Extracting semantically enriched events from biomedical literature. BMC Bioinforma 13(1):1–24
35. Ravikumar KE, Liu H, Cohn JD, Wall ME, Verspoor K (2011) Pattern learning through distant supervision for extraction of protein-residue associations in the biomedical literature, vol. 2. pp 59–65. IEEE, Honolulu. http://ieeexplore.ieee.org/document/6147049/.
36. Xia N, Lin H, Yang Z, Li Y (2011) Combining multiple disambiguation methods for gene mention normalization. Expert Syst Appl 38(7):7994–7999
37. Sarker A, Gonzalez G (2015) Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 53:196–207
38. Wu JL, Yu LC, Chang PC (2012) Detecting causality from online psychiatric texts using inter-sentential language patterns. BMC Med Inform Dec Making 12(1):1–10
39. Abacha AB, Zweigenbaum P (2011) A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. Lect Notes Comput Sci (Incl Subseries Lect Notes Artif Intell Lect Notes Bioinforma) 6609 LNCS(PART 2):139–150
40. Yu LC, Wu CH, Jang FL (2007) Psychiatric consultation record retrieval using scenario-based representation and multilevel mixture model. IEEE IEEE Trans Inf Technol Biomed 11(4):415–427
41. Musto C, Semeraro G, Lops P, Gemmis MD (2015) CrowdPulse: a framework for real-time semantic analysis of social streams. Inf Syst 54:127–146
42. García-Moya L, Kudama S, Aramburu MJ, Berlanga R (2013) Storing and analysing voice of the market data in the corporate data warehouse. Inf Syst Front 15(3):331–349
43. Eugenio BD, Green N, Subba R (2013) Detecting life events in feeds from Twitter. In: ICSC 2013: Proceedings of the IEEE Seventh International Conference on Semantic Computing. IEEE, Irvine, pp 274–277. http://ieeexplore.ieee.org/document/6693529/.
44. Torunoglu D, Telseren G, Sagturk O, Ganiz MC (2013) Wikipedia based semantic smoothing for twitter sentiment classification. In: INISTA 2013: Proceedings of the IEEE International Symposium on Innovations in Intelligent Systems and Applications. IEEE, Albena. pp 1–5
45. Cao Q, Duan W, Gan Q (2011) Exploring determinants of voting for the "helpfulness" of online user reviews: a text mining approach. Decis Support Syst 50(2):511–521
46. Levi A, Mokryn O, Diot C, Taft N (2012) Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In: RecSys'12: Proceedings of the sixth ACM Conference on Recommender Systems. ACM, New York. pp 115–122
47. He W, Shen J, Tian X, Li Y, Akula V, Yan G, et al (2015) Gaining competitive intelligence from social media data: evidence from two largest retail chains in the world. Ind Manag Data Syst 115(9):1622–1636
48. He W, Tian X, Chen Y, Chong D (2016) Actionable social media competitive analytics for understanding customer experiences. J Comput Inf Syst 56(2):145–155
49. Tian X, He W, Tao R, Akula V (2016) Mining online hotel reviews: a case study from hotels in China. In: AMCIS 2016: Proceedings of the 22nd Americas Conference on Information Systems. pp 1–8
50. ACM - Asian and Low-Resource Language Information Processing (TALLIP). http://tallip.acm.org/. Accessed 8 June 2016
51. Chen CL, Liu CL, Chang YC, Tsai HP (2011) Mining opinion holders and opinion patterns in US financial statements. In: TAAI 2011: Proceedings of the International Conference on Technologies and Applications of Artificial Intelligence. IEEE, Chung-Li, pp 62–68
52. Chen J, Liu J, Yu W, Wu P (2009) Combining lexical stability and improved lexical chain for unsupervised word sense disambiguation. In: KAM'09: Proceedings of the Second International Symposium on Knowledge Acquisition and Modeling. IEEE, Wuhan Vol. 1. pp 430–433. http://ieeexplore.ieee.org/document/5362135/.
53. Rusu D, Fortuna B, Grobelnik M, Mladenic D (2009) Semantic graphs derived from triplets with application in document summarization. Informatica (Slovenia) 33(3):357–362
54. Krachina O, Raskin V, Triezenberg K (2007) Reconciling privacy policies and regulations: ontological semantics perspective. In: Human Interface and the Management of Information. Interacting in Information Environments. Springer, Berlin, pp 730–739
55. Mansuy T, Hilderman RJ (2006) A characterization of WordNet features in Boolean models for text classification. In: AusDM 2006: Proceedings of the fifth Australasian Conference on Data Mining and Analystics. Australian Computer Society, Inc, Darlinghurst, pp 103–109
56. Ciaramita M, Gangemi A, Ratsch E, Šaric J, Rojas I (2005) Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In: IJCAI'05: Proceedings of the 19th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA. pp 659–664
57. Kim K, Chung BS, Choi Y, Lee S, Jung JY, Park J (2014) Language independent semantic kernels for short-text classification. Expert Syst Appl 41(2):735–743
58. Gujraniya D, Murty MN (2012) Efficient classification using phrases generated by topic models. In: ICPR 2012: Proceedings of the 21st International Conference on Pattern Recognition. IEEE, Tsukuba, pp 2331–2334
59. Du C, Zhuang F, He Q, Shi Z (2012) Multi-task semi-supervised semantic feature learning for classification. In: ICDM 2012: Proceedings of the IEEE 12th International Conference on Data Mining. IEEE, Brussels, pp 191–200. http://ieeexplore.ieee.org/document/6413903/.
60. Wu Q, Zhang C, Deng X, Jiang C (2011) LDA-based model for topic evolution mining on text. In: ICCSE 2011: Proceedings of the 6th International Conference on Computer Science & Education. IEEE, Singapore, pp 946–949
61. Lu X, Zheng B, Velivelli A, Zhai C (2006) Enhancing text categorization with semantic-enriched representation and training data augmentation. J Am Med Inform Assoc 13(5):526–535
62. Wu J, Dang Y, Pan D, Xuan Z, Liu Q (2010) Textual knowledge representation through the semantic-based graph structure in

clustering applications. In: HICSS 2010: Proceedings of the 43rd Hawaii International Conference on System Sciences. IEEE, Washington, pp 1–8

63. Princeton University - WordNet. http://wordnet.princeton.edu/. Accessed 8 June 2016
64. Fellbaum C (1998) WordNet: an electronic lexical database. MIT Press, Cambridge
65. Weller K (2010) Knowledge representation in the social semantic web. Walter de Gruyter
66. Weller K, et al (2007) Folksonomies and ontologies: two new players in indexing and knowledge representation. In: Proceedings of the Online Information Conference. pp 108–115
67. Wei TA, Lu YC, Chang HB, Zhou QA, Bao XD (2015) A semantic approach for text clustering using WordNet and lexical chains. Expert Syst Appl 42(4):2264–2275
68. Li J, Zhao Y, Liu B (2009) Fully automatic text categorization by exploiting wordnet. In: Information Retrieval Technology. Springer, Berlin, pp 1–12
69. Mansuy TN, Hilderman RJ (2006) Evaluating WordNet features in text classification models. In: FLAIRS Conference 2006: Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference. AAAI PRESS, Florida, pp 568–573
70. Shin Y, Ahn Y, Kim H, Lee SG (2015) Exploiting synonymy to measure semantic similarity of sentences. In: IMCOM '15: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. ACM, New York, pp 40:1–40:4
71. Batet M, Valls A, Gibert K (2010) Performance of ontology-based semantic similarities in clustering. In: Artificial Intelligence and Soft Computing. Springer, Berlin, pp 281–288
72. Basu S, Mooney RJ, Pasupuleti KV, Ghosh J (2001) Evaluating the novelty of text-mined rules using lexical knowledge. In: KDD'01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, San Francisco, pp 233–238
73. Wikipedia. https://www.wikipedia.org/. Accessed 8 June 2016
74. Kim HJA, Hong KJA, Chang JYb (2015) Semantically enriching text representation model for document clustering. In: Proceedings of the ACM Symposium on Applied Computing, ACM, New York, pp 922–925. http://dl.acm.org.ez67.periodicos.capes.gov.br/citation.cfm?id=2696055.
75. Yun J, Jing L, Yu J, Huang H (2011) Unsupervised feature weighting based on local feature relatedness. In: Advances in Knowledge Discovery and Data Mining. Springer, Berlin, pp 38–49
76. Gabrilovich E, Markovitch S (2009) Wikipedia-based semantic interpretation for natural language processing. J Artif Intell Res 34:443–498
77. Hu X, Zhang X, Lu C, Park EK, Zhou X (2009) Exploiting Wikipedia as external knowledge for document clustering. In: KDD'09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, pp 389–396
78. Mizzaro S, Pavan M, Scagnetto I, Valenti M (2014) Short text categorization exploiting contextual enrichment and external knowledge. In: Proceedings of the First International Workshop on Social Media Retrieval and Analysis. ACM, New York, pp 57–62
79. Janik M, Kochut KJ (2008) Wikipedia in action: ontological knowledge in text categorization. In: ICSC 2008: Proceedings of the International Conference on Semantic Computing. IEEE, Santa Monica, pp 268–275
80. Chang MW, Ratinov LA, Roth D, Srikumar V (2008) Importance of semantic representation: dataless classification. In: AAAI-08: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. pp 830–835
81. Medelyan O, Milne D, Legg C, Witten IH (2009) Mining meaning from Wikipedia. Int J Human-Computer Stud 67(9):716–754
82. HowNet Knowledge Database. http://www.keenage.com/. Accessed 8 June 2016
83. Jin CX, Zhou HY, Bai QC (2012) Short text clustering algorithm with feature keyword expansion. Adv Mater Res 532:1716–1720
84. Liu Z, Yu W, Chen W, Wang S, Wu F (2010) Short text feature selection for micro-blog mining. In: CiSE 2010: Proceedings of the International Conference on Computational Intelligence and Software Engineering. IEEE, Wuhan, pp 1–4
85. Hu P, He T, Ji D, Wang M (2004) A study of Chinese text summarization using adaptive clustering of paragraphs. In: CIT'04: Proceedings of the Fourth International Conference on Computer and Information Technology. IEEE, Wuhan, pp 1159–1164
86. Zhu ZY, Dong SJ, Yu CL, He J (2011) A text hybrid clustering algorithm based on HowNet semantics. Key Eng Mater 474:2071–2078
87. Zheng D, Liu H, Zhao T (2011) Search results clustering based on a linear weighting method of similarity. In: IALP 2011: Proceedings of the International Conference on Asian Language Processing. IEEE, Penang, pp 123–126
88. Wang R (2010) Cognitive-based emotion classifier of Chinese vocabulary design. In: ISISE 2010: Proceedings of the International Symposium on Information Science and Engineering. IEEE. pp 582–585
89. Thorleuchter D, Van den Poel D (2014) Semantic compared cross impact analysis. Expert Syst Appl 41(7):3477–3483
90. Roussinov D, Turetken O (2009) Exploring models for semantic category verification. Inf Syst 34(8):753–765
91. Zelikovitz S, Kogan M (2006) Using Web searches on important words to create background sets for LSI classification. In: FLAIRS Conference 2006: Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference Vol. 1. pp 298–603
92. SentiWordNet. http://sentiwordnet.isti.cnr.it/. Accessed 8 June 2016
93. Al Nasseri A, Tucker A, de Cesare S (2015) Quantifying StockTwits semantic terms' trading behavior in financial markets: an effective application of decision tree algorithms. Expert Syst Appl 42(23):9192–9210
94. Kumar V, Minz S (2013) Mood classifiaction of lyrics using SentiWordNet. In: ICCCI 2013: Proceedings of the International Conference on Computer Communication and Informatics. IEEE, Coimbatore, pp 1–5
95. Unified Medical Language System (UMLS) Metathesaurus. https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/. Accessed 8 June 2016
96. Garla VN, Brandt C (2012) Ontology-guided feature engineering for clinical text classification. J Biomed Inform 45(5):992–998
97. Plaza L, Díaz A, Gervás P (2011) A semantic graph-based approach to biomedical summarisation. Artif Intell Med 53(1):1–14
98. Aljaber B, Martinez D, Stokes N, Bailey J (2011) Improving MeSH classification of biomedical articles using citation contexts. J Biomed Inform 44(5):881–896
99. Medical Subject Headings (MeSH). https://www.nlm.nih.gov/mesh/. Accessed 8 June 2016
100. Logeswari S, Premalatha K (2013) Biomedical document clustering using ontology based concept weight. In: ICCCI 2013: Proceedings of the International Conference on Computer Communication and Informatics. IEEE, Coimbatore, pp 1–4
101. Nguyen SH, Jaśkiewicz G, Świeboda W, Nguyen HS (2012) Enhancing search result clustering with semantic indexing. In: SoICT'12: Proceedings of the Third Symposium on Information and Communication Technology. ACM, New York, pp 71–80
102. Ginter F, Pyysalo S, Boberg J, Järvinen J, Salakoski T (2004) Ontology-based feature transformations: a data-driven approach. In: Advances in Natural Language Processing. Springer, Berlin, pp 279–290
103. Kanavos A, Makris C, Theodoridis E (2012) On topic categorization of PubMed query results. In: Artificial Intelligence Applications and Innovations. Springer. pp 556–565
104. Zheng HT, Borchert C, Kim HG (2008) Exploiting gene ontology to conceptualize biomedical document collections. In: The Semantic Web. Springer, Berlin, pp 375–389
105. Jin B, Muller B, Zhai C, Lu X (2008) Multi-label literature classification based on the Gene Ontology graph. BMC Bioinforma 9(1):525
106. Mannai M, Ben Abdessalem Karaa W (2013) Bayesian information extraction network for Medline abstract. In: 2013 World Congress on Computer and Information Technology (WCCIT). IEEE, Sousse, pp 1-3.
107. Jiana B, Tingyu L, Tianfang Y (2012) Event information extraction approach based on complex Chinese texts. In: IALP 2012: Proceedings of the International Conference on Asian Language Processing. pp 61–64
108. Hengliang W, Weiwei Z (2012) A web information extraction method based on ontology. Adv Inf Sci Serv Sci 4(8):199–206
109. Aghassi H, Sheykhlar Z (2012) Extending information retrieval by adjusting text feature vectors. Commun Comput Inform Sci 295 CCIS:133–142
110. Bharathi G, Venkatesan D (2012) Improving information retrieval using document clusters and semantic synonym extraction. J Theor Appl Inf Technol 36(2):167–173

111. Egozi O, Markovitch S, Gabrilovich E (2011) Concept-based information retrieval using explicit semantic analysis. ACM Trans Inf Syst 29(2):8:1–8:34
112. Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL (2015) Text mining of news-headlines for FOREX market prediction: a multi-layer dimension reduction algorithm with semantics and sentiment. Expert Syst Appl 42(1):306–324
113. Batool R, Khattak AM, Maqbool J, Lee S (2013) Precise tweet classification and sentiment analysis. In: 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS). IEEE, Niigata, pp 461–466
114. Veselovská K (2012) Sentence-level sentiment analysis in Czech. In: WIMS'12:Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics. ACM, New York, pp 65:1–65:4
115. Petersen MK, Hansen LK (2012) On an emotional node: modeling sentiment in graphs of action verbs. In: 2012 International Conference on Audio, Language and Image Processing. IEEE, Shanghai, pp 308–313
116. Domínguez García R, Schmidt S, Rensing C, Steinmetz R (2012) Automatic taxonomy extraction in different languages using wikipedia and minimal language-specific information. Lect Notes Comp Sci (Incl Subseries Lect Notes Artif Intell Lect Notes Bioinforma) 7181 LNCS (PART 1):42–53
117. Punuru J, Chen J (2012) Learning non-taxonomical semantic relations from domain texts. J Intell Inf Syst 38(1):191–207
118. Stenetorp P, Soyer H, Pyysalo S, Ananiadou S, Chikayama T (2012) Size (and domain) matters: evaluating semantic word space representations for biomedical text. In: SMBM 2012: Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine, pp 42–49
119. Froud H, Lachkar A, Ouatik SA (2012) Stemming versus light stemming for measuring the simitilarity between Arabic words with latent semantic analysis model. In: 2012 Colloquium in Information Science and Technology. IEEE, Fez, pp 69–73
120. Kuhn A, Ducasse S, Gírba T (2007) Semantic clustering: identifying topics in source code. Inf Softw Technol 49:230–243
121. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
122. Zrigui M, Ayadi R, Mars M, Maraoui M (2012) Arabic text classification framework based on latent dirichlet allocation. J Comput Inf Technol 20(2):125–140
123. Liu Z, Li M, Liu Y, Ponraj M (2011) Performance evaluation of latent Dirichlet allocation in text mining. In: FSKD 2011: Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, Shanghai Vol. 4. pp 2695–2698
124. Xiang W, Yan J, Ruhua C, Hua F (2013) Improving text categorization with semantic knowledge in Wikipedia. IEICE Trans Inf Syst 96(12):2786–2794
125. Spanakis G, Siolas G, Stafylopatis A (2012) Exploiting Wikipedia knowledge for conceptual hierarchical clustering of documents. Comput J 55(3):299–312
126. Andreasen T, Bulskov H, Jensen PA, Lassen T (2011) Extracting conceptual feature structures from text. In: ISMIS 2011: Proceedings 19th International Symposium on Methodologies for Intelligent Systems. Springer, Berlin, pp 396–406
127. Goossen F, IJntema W, Frasincar F, Hogenboom F, Kaymak U (2011) News personalization using the CF-IDF semantic recommender. In: WIMS'11: Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, New York, p 10
128. Huang A, Milne D, Frank E, Witten IH (2008) Clustering documents with active learning using Wikipedia. In: ICDM'08: Eighth IEEE International Conference on Data Mining. IEEE, Pisa, pp 839–844
129. Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI-07: Proceedings of the 20th International Joint Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc, San Francisco, pp 1606–1611. http://dl.acm.org.ez67.periodicos.capes.gov.br/citation.cfm?id=1625535.
130. Navigli R, Faralli S, Soroa A, de Lacalle O, Agirre E (2011) Two birds with one stone: learning semantic models for text Categorization and word sense disambiguation. In: CIKM'11: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, Glasgow, pp 2317–2320
131. Mostafa MS, Haggag MH, Gomaa WH (2008) Document clustering using word sense disambiguation. In: SEDE 2008: Proceedings of 17th International Conference on Software Engineering and Data Engineering. pp 19–24
132. Andreopoulos B, Alexopoulou D, Schroeder M (2008) Word sense disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering. Int J Data Min Bioinforma 2(3):193–215
133. Koeling R, McCarthy D, Carroll J (2007) Text categorization for improved priors of word meaning. In: Computational Linguistics and Intelligent Text Processing. Springer, Berlin, pp 241–252
134. Sharma A, Swaminathan R, Yang H (2010) A verb-centric approach for relationship extraction in biomedical text. In: ICSC 2010: Proceedings of the IEEE Fourth International Conference on Semantic Computing. IEEE, Pittsburgh, pp 377–385
135. Wang W, Zhao D, Zou L, Wang D, Zheng W (2010) Extracting 5W1H event semantic elements from Chinese online news. In: WAIM 2010: Proceedings of the Workshops of the 11th International Conference on Web-Age Information Management. Springer, Berlin, pp 644–655
136. Rebholz-Schuhmann D, Jimeno-Yepes A, Arregui M, Kirsch H (2010) Measuring prediction capacity of individual verbs for the identification of protein interactions. J Biomed Inform 43(2):200–207
137. Van Der Horn P, Bakker B, Geleijnse G, Korst J, Kurkin S (2008) Classifying verbs in biomedical text using subject-verb-object relationships. In: SMBM 2008: Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine. pp 137–140
138. Kontos J, Malagardi I, Alexandris C, Bouligaraki M (2000) Greek verb semantic processing for stock market text mining. In: NLP'00: Proceedings of the Second International Conference on Natural Language Processing. Springer-Verlag, London. pp 395–405
139. Stankov I, Todorov D, Setchi R (2013) Enhanced cross-domain document clustering with a semantically enhanced text stemmer (SETS). Int J Knowl-Based Intell Eng Syst 17(2):113–126
140. Huang CH, Yin J, Hou F (2011) A text similarity measurement combining word semantic information with TF-IDF method. Jisuanji Xuebao(Chin J Comput) 34(5):856–864
141. Doan S, Kawazoe A, Conway M, Collier N (2009) Towards role-based filtering of disease outbreak reports. J Biomed Inform 42(5):773–780
142. Meng X, Chen Q, Wang X (2008) Semantic feature reduction in chinese document clustering. In: SMC 2008: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. IEEE, Singapore, pp 3721–3726
143. Freitas A, O'Riain S, Curry E, da Silva JCP, Carvalho DS (2013) Representing texts as contextualized entity-centric linked data graphs. In: DEXA 2013: Proceedings of the 24th International Workshop on Database and Expert Systems Applications. IEEE, Los Alamitos, pp 133–137
144. Fathy I, Fadl D, Aref M (2012) Rich semantic representation based approach for text generation. In: INFOS 2012: Proceedings of the 8th International Conference on Informatics and Systems. IEEE, Cairo, pp NLP–20
145. Wu J, Xuan Z, Pan D (2011) Enhancing text representation for classification tasks with semantic graph structures. Int J Innov Comput Inf Control (ICIC) 7(5):2689–2698
146. Alencar ROD, Davis Jr CA, Gonçalves MA (2010) Geographical classification of documents using evidence from Wikipedia. In: GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval. ACM, New York, p 12
147. Smirnov I, Tikhomirov I (2009) Heterogeneous semantic networks for text representation in intelligent search engine EXACTUS. In: SENSE'09: Proceedings of the Workshop on Conceptual Structures for Extracting Natural Language Semantics. pp 1–9
148. Chau R, Tsoi AC, Hagenbuchner M, Lee V (2009) A conceptlink graph for text structure mining. In: ACSC'09: Proceedings of the Thirty-Second Australasian Conference on Computer Science - Volume 91. Australian Computer Society, Inc., Darlinghurst, pp 141–150
149. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117
150. Lebret R, Collobert R (2015) Rehabilitation of count-based models for word vector representations. Lect Notes Comput Sci (Incl Subseries Lect Notes Artif Intell Lect Notes Bioinforma) 9041:417–429
151. Li R, Shindo H (2015) Distributed document representation for document classification. Lect Notes Comput Sci (Incl Subseries Lect Notes Artif Intell Lect Notes Bioinforma) 9077:212–225

152. Sohrab MG, Miwa M, Sasaki Y (2015) Centroid-means-embedding: an approach to infusing word embeddings into features for text classification. Lect Notes Comput Sci (Incl Subseries Lect Notes Artif Intell Lect Notes Bioinforma) 9077:289–300
153. Wang P, Xu B, Xu J, Tian G, Liu CL, Hao H (2016) Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing 174:806–814
154. Zhang C, Zhang L, Wang CJ, Xie JY (2014) Text summarization based on sentence selection with semantic representation. In: Proceedings of the International Conference on Tools with Artificial Intelligence, Vol. 2014-December. IEEE, Limassol. 584–590
155. Vulić I, Moens MF (2015) Monolingual and cross-lingual information retrieval models based on (Bilingual) word embeddings. In: SIGIR'15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, pp 363–372
156. Kamal A, Abulaish M, Anwar T (2012) Mining feature-opinion pairs and their reliability scores from web opinion sources. In: WIMS'12: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. ACM, New York, p 15
157. Kong L, Yan R, He Y, Zhang Y, Zhang Z, Fu L (2011) DVD: a model for event diversified versions discovery. In: Web Technologies and Applications. Springer, Berlin, pp 168–180
158. Jing L, Yun J, Yu J, Huang J (2011) High-order co-clustering text data on semantics-based representation model. In: Advances in Knowledge Discovery and Data Mining. Springer, Berlin, pp 171–182
159. Krajewski R, Rybinski H, Kozlowski M (2016) A novel method for dictionary translation. J Intell Inf Syst 47(3):491–514
160. Luo Z, Miotto R, Weng C (2013) A human–computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. J Biomed Inform 46(1):33–39
161. Kayed A (2005) Building e-laws ontology: new approach. In: Proceedings of the On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops. Springer, Berlin, pp 826–835
162. Sevenster M, van Ommering R, Qian Y (2012) Automatically correlating clinical findings and body locations in radiology reports using MedLEE. J Digit Imaging 25(2):240–249
163. Volkova S, Caragea D, Hsu WH, Drouhard J, Fowles L (2010) Boosting biomedical entity extraction by using syntactic patterns for semantic relation discovery. In: WI-IAT 2010: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. IEEE, Toronto Vol. 1. pp 272–278
164. Waltinger U, Mehler A (2009) Social semantics and its evaluation by means of semantic relatedness and open topic models. In: WI-IAT'09: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, Milan, pp 42–49
165. Kass A, Cowell-Shah C (2006) Using lightweight NLP and semantic modeling to realize the internet's potential as a corporate radar. In: AAAI Fall Symposium. AAAI PRESS
166. Blake C (2010) Beyond genes, proteins, and abstracts: identifying scientific claims from full-text biomedical articles. J Biomed Inform 43(2):173–189
167. Hu J, Fang L, Cao Y, Zeng HJ, Li H, Yang Q, et al (2008) Enhancing text clustering by leveraging Wikipedia semantics. In: SIGIR'08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, pp 179–186
168. Lu CY, Lin SH, Liu JC, Cruz-Lara S, Hong JS (2010) Automatic event-level textual emotion sensing using mutual action histogram between entities. Expert Syst Appl 37(2):1643–1653
169. Ahmed ST, Nair R, Patel C, Davulcu H (2009) BioEve: bio-molecular event extraction from text using semantic classification and dependency parsing. In: BioNLP'09: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics. pp 99–102
170. Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recogn Lett 31(8):651–666
171. Wordle. http://www.wordle.net/. Accessed 15 June 2016