Received: 4 January 2018 | Revised: 13 June 2018 | Accepted: 25 December 2018

DOI: 10.1002/widm.1302

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

ADVANCED REVIEW

An advanced review on text mining in medicine

Carmen Luque[1] | José M. Luna[1,3] | Maria Luque[3] | Sebastian Ventura[1,2,3]

[1]Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory, Maimonides Biomedical Research Institute of Cordoba, Cordoba, Spain

[2]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[3]Department of Computer Science and Numerical Analysis, University of Cordoba, Cordoba, Spain

Correspondence
Sebastian Ventura, Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory, Maimonides Biomedical Research Institute of Cordoba, Córdoba, Spain.
Email: sventura@uco.es

Funding information
European Regional Development Fund, Grant/Award Number: TIN2017-83445-P; Spanish Ministry of Economy and Competitiveness

Health care professionals produce abundant textual information in their daily clinical practice and this information is stored in many diverse sources and, generally, in textual form. The extraction of insights from all the gathered information, mainly unstructured and lacking normalization, is one of the major challenges in computational medicine. In this respect, text mining (TM) assembles different techniques to derive valuable insights from unstructured textual data so it has led to be especially relevant in medicine. The aim of this paper is therefore to provide an extensive review of existing techniques and resources to perform TM tasks in medicine. In this review, more than 90 relevant research studies have been analyzed, describing the most important practical applications, terminological resources, tools, and open challenges of TM in medicine.

This article is categorized under:
Application Areas > Health Care
Algorithmic Development > Biological Data Mining
Algorithmic Development > Hierarchies and Trees
Algorithmic Development > Model Combining

KEYWORDS
medicine, text mining, text mining tools

1 | INTRODUCTION

Over the last decades, the quantity of available information daily produced in medicine is growing considerably with a special emphasis on that generated by health care professionals in their general daily practices (Feldman, Hazekamp, & Chawla, 2016). The health of the patients is regularly described by thousands of doctors; the results come in the form of textual information that is stored in different format files such as clinical records, discharge summaries, clinical monitoring sheets, or radiological reports. As a consequence of these unstructured textual data sources, the extraction of useful knowledge for decision-making and the reusability of such information is hampered.

Currently, the main problem to be faced by any health care professional is not simply obtaining any available clinical information from databases but a promising subset including the most relevant and useful information. The final aim is therefore to transform this information into knowledge so professionals in the field might leverage their daily practice. Nevertheless, this is not a trivial task since clinical information is very different from any other and it usually includes some special features: high ambiguity and complex vocabulary; absence of terminological standardization; short sentences that may contain grammatical errors; overuse of acronyms; structured and unstructured data are usually combined; and texts are normally written in a narrative form.

The discovery of hidden knowledge in this amount of unstructured information is essential to provide support on the decision-taking process that is carried out by the professionals every day. In this regard, the term text mining (TM) gathers the most useful techniques to derive high-quality structured information from unstructured textual data (Feldman & Sanger,

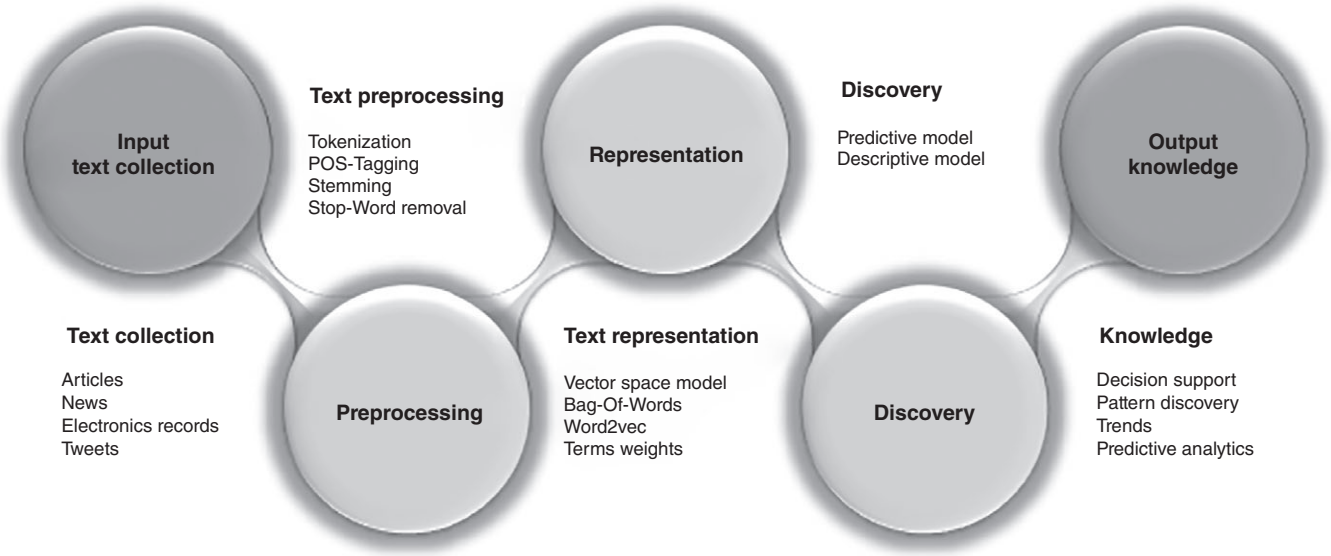

FIGURE 1 Text mining process

2007). It is a process in which useful and unknown knowledge is extracted from textual data by applying a series of phases (see Figure 1): (a) Preprocessing stage, where textual and unstructured input data are standardized and cleaned by means of different natural language processing (NLP) techniques (Collobert et al., 2011), for example, tokenization and stemming; (b) Text representation stage, where the unstructured input data is transformed into a suitable representation model that allows an efficient analysis to be performed in subsequent phases, for example, Bag-Of-Words (BOW) (Zhang, Jin, & Zhou, 2010); (c) Discovery stage, where useful, unexpected, and unknown information is extracted from textual data collections through the application of certain methods and techniques (Allahyari et al., 2017), for example, classification, clustering, etc. Due to space limitation, a more detailed description about TM and the aforementioned phases is available at http://www.uco.es/kdis/textminingmedicine/#tm. The applicability of these techniques to medicine is keystone to ease the labor of health care professionals in both research and clinical daily issues, for example, the prediction of a specific disease according to the features of each patient or the development of systems to support medical diagnostic decision-making.

TM techniques (Gupta, Lehal, et al., 2009) have been widely studied and analyzed from a technical perspective. Nevertheless, due to the increasing applicability of TM to medicine (Feldman & Sanger, 2007), it becomes essential to provide a general analysis of existing approaches and methodologies that have contributed to improve the health care. In this regard, the aim of this paper is to review the most interesting research works published on the use of TM (Feldman & Sanger, 2007) in medicine. More than 90 research articles have been analyzed with special attention paid to those involving important applications such as named entity recognition (NER), relationship extraction, text summarization, and terminology extraction, among others. The final aim is therefore to provide the readers with an extensive revision about existing techniques and resources to perform TM in medicine.

The paper is structured as follows. Section 2 describes the main contributions of TM to medicine. Some terminological resources and tools used in textual analysis are also described in Section 3. Finally, Section 4 summarizes some open challenges, while Sections 5 and 6 present the lesson learned and some concluding remarks.

2 | TM APPLIED TO MEDICINE

Since the beginning of the century, when first research works appeared, TM techniques (Gupta et al., 2009) applied to health care tasks (diagnosis, treatment, and prevention of different diseases) are denoting an increasing interest (Chen, Fuller, Friedman, & Hersh, 2005), as it is demonstrated by the growth number of articles published in this regard on Embase and PubMed databases (see Figure 2). The number of both biological and clinical challenges in biomedical TM has increased in the last years (in average, more than four different challenges take per year from 2008), which have produced an increment of research articles in the matter (Huang & Lu, 2016). TM techniques have been considered on multiple research studies in medicine (see Table 1), the most important ones are described in this section. A detailed summary table of these research studies is also provided, including the application area, the used techniques, and the obtained insights. Due to space limitation, the summary table is available at http://www.uco.es/kdis/textminingmedicine/#summarytable.

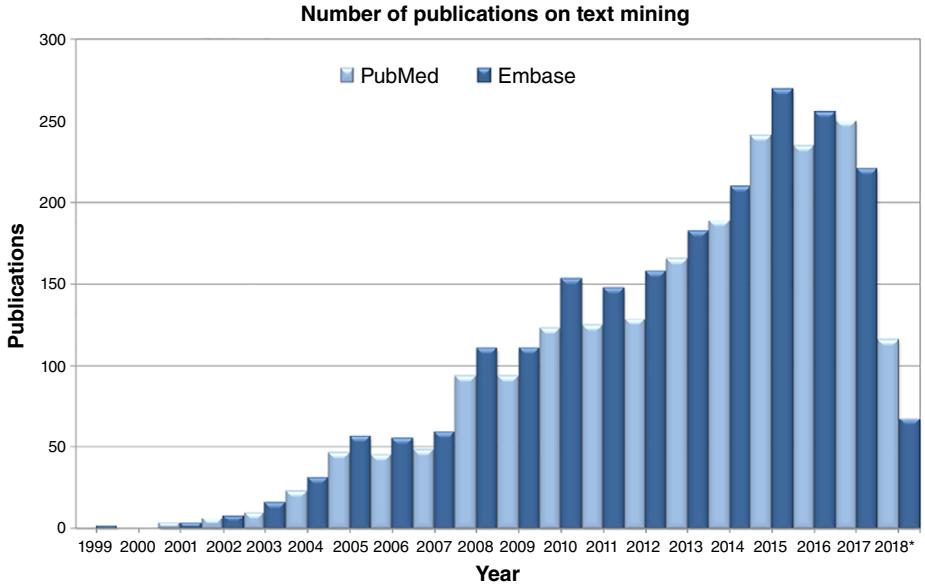

FIGURE 2 Number of scientific publications returned by PubMed and Embase biomedical databases (keywords: text mining. Last updated June 2018)

TABLE 1 Summary of references where text mining was applied to medicine

Applications	References
Named entity recognition	Benton et al. (2011), Carrero, Cortizo, and Gomez (2008), Ferrandez, South, Shen, and Meystre (2012), Kipper-Schuler, Kaggal, Masanz, Ogren, and Savova (2008), Lin, Chen, and Brown (2013), Rink, Harabagiu, and Roberts (2011), Roberts and Harabagiu (2011), Roberts, Rink, and Harabagiu (2013), Skeppstedt, Kvist, Nilsson, and Dalianis (2014), Uzuner, Mailoa, Ryan, and Sibanda (2010), Wang and Patrick (2009), Xia et al. (2013), Zhu, Cherry, Kiritchenko, Martin, and De Bruijn (2013)
Hypothesis generation and knowledge discovery	Baron et al. (2013), Byrd, Steinhubl, Sun, Ebadollahi, and Stewart (2014), Cole et al. (2013), Collier (2012), Heintzelman et al. (2013), Leeper et al. (2013), Tafti et al. (2017), and Yang et al. (2017)
Text summarization	Afantenos, Karkaletsis, and Stamatopoulos (2005), Elhadad, Kan, Klavans, and McKeown (2005), Fiszman, Demner-Fushman, Kilicoglu, and Rindflesch (2009), Rindflesch, Kilicoglu, Fiszman, Rosemblat, and Shin (2011), Sarkar, Nasipuri, and Ghose (2011), Zhang et al. (2011)
Terminology extraction	Fabian, Wachter, and Schroeder (2012), Fang, Huang, Chen, and Juan (2008), Luther et al. (2011), Roberts et al. (2009), Van Mulligen et al. (2012), and Xie, Ding, Han, and Wu (2013)
Text classification	Asghar et al. (2013), Castro et al. (2015), Frunza, Inkpen, Matwin, Klement, and O'blenis (2011), Goldstein, Arzumtsyan, and Uzuner (2007), Harpaz et al. (2014), Jonnagaddala, Dai, Ray, and Liaw (2015), Metais, Nakache, and Timsit (2006), Pakhomov et al. (2007), Vijayakrishnan et al. (2014), Yetisgen-Yildiz and Pratt (2005), and Zuccon et al. (2013)

2.1 | Named entity recognition

A named entity can be defined as a word (or set of words) that identifies a person, an organization, a place, a date, a specific time, a percentage or a quantity. NER is, therefore, the process of discovering named entities in different chunks of textual data (Nadeau & Sekine, 2007). In the medical field NER systems (Wang & Patrick, 2009) are mainly used to extract concepts as well as terms (the name of a disease, a symptom, or a concrete anatomical region from a set of clinical reports) that are somehow disperse among several textual sources due to the natural language. There are basically four types of approaches to deal with NER in the medical field (Sun, Cai, Liu, Fang, & Wang, 2017): rule-based, dictionary-based, machine learning (ML)-based (Michalski, Carbonell, & Mitchell, 2013), and a hybridization of the aforementioned approaches. To demonstrate the importance of NER in medicine, a wide number of research works are analyzed. First, the focus is on the extraction of medical concepts from clinical reports written in different languages. Second, the discovery of concepts related to temporal expressions is analyzed. Third, the attention is paid on personal data anonymization. Finally, the extraction of relationships between entities is analyzed.

2.1.1 | Medical concepts from clinical reports

Roberts and Harabagiu (Roberts & Harabagiu, 2011) proposed the extraction of medical concepts (e.g., disease, drug, injury) in clinical texts under a ML-based approach (Michalski et al., 2013). Authors presented a classification task by considering two ML approaches: Support vector machine (SVM) (Yu, Liu, Valdez, Gwinn, & Khoury, 2010) and conditional random fields (CRF) (Lafferty, McCallum, & Pereira, 2001). Additionally, the NegEx (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001) algorithm was used to detect negations. For concept extraction, several external sources of knowledge were

used including MetaMap (Aronson & Lang, 2010), Wikipedia, and WordNet. The resulting model was evaluated according to the 2010 i2b2/VA challenge data (Workshop on NLP Challenges for Clinical Records), obtaining a *F*-measure value of 0.796 for the concept extraction task (the best i2b2 submission value was 0.852, and the median value for all the submissions was 0.778). Kipper-Schuler et al. (2008) analyzed and evaluated the NER component within the information extraction (IE) system of a specific clinic, which aimed at discovering medical entities such as diseases, symptoms, medications, procedures, etc. In this work authors proposed a dictionary-based approach by using the unified medical language system (UMLS) metathesaurus (Bodenreider, 2004), which was also expanded with synonyms. Authors compared the dictionary look-up model to different ML algorithms (Michalski et al., 2013) (CRF, Lafferty et al., 2001; and SVM, Yu et al., 2010). Results showed that CRF with multiple features significantly outperformed a single feature of dictionary look-up (baseline system), obtaining the highest performance with a value of 0.860 in *F*-measure. Xia et al. (2013) addressed the task of disease recognition from clinical texts that was proposed in the clinical e-science framework (CLEF) eHealth 2013 conference. Authors used a dictionary-based approach, combining MetaMap (Aronson & Lang, 2010) and cTAKES (Savova et al., 2010) to solve the NER task as well as the normalization of disorders. Results showed that the combination of these two systems outperformed MetaMap and cTAKES in isolation, obtaining an improvement around 4% in *F*-measure.

Finally, it should be highlighted that, even when most of the current NER systems have a really good performance on English texts (Dandapat & Way, 2016), some authors have started to explore their application on different languages. Skeppstedt et al. (2014) analyzed the performance on Swedish texts for the extraction of four types of medical entities (disorder, finding, pharmaceutical drug, and body structure) through a ML-based approach. Here, authors considered inside-outside-beginning encoding (Ramshaw & Marcus, 1999) for annotated entities and the CRF (Lafferty et al., 2001) algorithm, obtaining similar results (a *F*-measure value of 0.810 for disorder recognition) to those obtained on English texts. Carrero et al. (Carrero et al., 2008) proposed a cross-lingual system, called GALEN, based on dictionaries to retrieve cross-lingual information related to medical records. Authors combined MetaMap Transfer tool (MMTx) (Meystre & Haug, 2005) and automatic translation techniques, to extract named entities from Spanish texts. They also evaluated the proposed system, reaching to the conclusion that the results are similar to those obtained on English texts (average similarities of 79.42%).

2.1.2 | Time expressions

The accurate extraction of time expressions in the medical field (date of onset of a disease, duration and frequency of the treatment, among others) is a really important and arduous task that have been widely studied. Lin et al. (2013) presented the MedTime system, a temporal information extraction system including differentt rule-based and ML (Michalski et al., 2013) procedures (SVM, Yu et al., 2010; and CRF, Lafferty et al., 2001). One of the objectives of this hybrid system was the extraction, in clinical texts, of entities related to temporal expressions such as date, time, duration, and frequency. Among others, the authors considered MetaMap (Aronson & Lang, 2010) for extracting features of medical lexicon and semantic types, and Mallet (McCallum, 2002) for temporal expressions annotation. Authors demonstrated the efficiency of the hybrid approach, obtaining a *F*-measure value of 0.879 (the best submission value in the i2b2 challenge was 0.917, whereas the median value of all the submissions was 0.792). Roberts et al. (2013) proposed a hybrid system for automatic recognition of events, temporal expressions, and temporal relations in clinical records. Authors combined different ML methods (Michalski et al., 2013) and a rule-based method. As a result, 0.893 for *F*-measure and 0.548 for the task of extracting temporary expressions were obtained on 2012 i2b2 challenge.

2.1.3 | Personal data anonymization

The protection of personal data has become a challenge for many health institutions, especially with the rising computerization of almost any clinical record. Ferrandez et al. (2012) applied different anonymization techniques based on NER to remove or disguise sensitive information. Authors presented a hybrid system based on either rule-based, dictionary and ML (Michalski et al., 2013) methodologies with the aim of improving the person names de-identification task. Authors compared their system with five existing systems in the field of entity extraction and de-identification. Results demonstrated an improvement in more than a 26% for the F2-measure metric when the proposal is compared to the best system. Benton et al. (2011) presented a proposal to remove telephone numbers, names, e-mail addresses, and other identifying data from medical message boards. To carry out this task, authors used a hybrid approach, based on rules and ML (Michalski et al., 2013). A CRF (Lafferty et al., 2001) model was used to tag the identifiers, whereas two corpora (one based on breast cancer and another on arthritis) were considered to train and validate the model. Authors evaluated their system against a well-known de-identification system, achieving a good performance in the recall evaluation measure (0.981 vs. 0.730).

2.1.4 | Relationship extraction

The discovery of semantic relationships, for example, connections among diseases and symptoms, between medical concepts is essential. Uzuner et al. (2010) discovered different relationships among several clinic entities in a set of hospital discharge reports. Based on the semantic types defined in the UMLS Meta-thesaurus (Bodenreider, 2004), the following relations were analyzed: disease–treatment, disease–test, and disease–symptom. Authors used a ML-based approach that included a SVM (Yu et al., 2010). For each pair of concepts that were included in a sentence, their relationships were determined by a semantic relation classifier based on SVM. In order to evaluate the proposal, two baseline systems were considered, obtaining a *F*-measure value that was 15% better than the one obtained by the best baseline system. Rink et al. (2011) used some electronic medical records to analyze eight associations between medical problems, treatments, and tests. To carry out this task, authors used a ML-based approach (CRF, Lafferty et al., 2001; SVM, Yu et al., 2010). They used CRF (Lafferty et al., 2001) to extract medical concepts, and different SVM (Yu et al., 2010) models to identify the relationships between the extracted concepts. Results were improved by considering some external resources such as Wikipedia or WordNet and NLP tools for concept discovery features. The proposal obtained the best performance in *F*-measure (a value of 0.736) among all the approaches presented in the 2010 i2b2 NLP Challenge. Zhu et al. (Zhu et al., 2013) described a model to identify semantic relations among medical concepts (problems, tests, and treatments) from real-world discharge summaries and progress reports. Three types of relations were analyzed: treatment–problem, test–problem, and problem–problem. To carry out this task, authors used a hybrid approach, based on ML (Michalski et al., 2013), dictionaries, and rules. Medical concepts were extracted by using a concept-recognition system (a discriminative semi-Markov model), whereas the relationships between concepts were obtained by different classification approaches (Dreiseitl & Ohno-Machado, 2002) (e.g., SVM, kNN, logistic regression). Results showed the importance of performing a good feature selection process through different external sources of knowledge, achieving an improvement in *F*-measure (a value of 0.742 was obtained) for the extraction of relationships between medical entities.

2.2 | Hypotheses generation and knowledge discovery

The discovery of new hypotheses and hidden knowledge on textual data is essential to provide health care professionals with important insights that can be used in their daily practices, and it also supports their research works. To this aim, the application of TM techniques in medicine is essential to support the discovery of such valuable knowledge that is useful to detect risk factors, symptoms, and critical events of a patient, and facilitate, therefore, the arduous task of decision-making that is daily carried out by health professionals.

2.2.1 | Hypotheses generation

Baron et al. (2013) performed a meta-analysis to identify adverse effects of aspirin usage, considering TM (Gupta et al., 2009) as an automated procedure to score articles for potential relevance to a posteriori meta-analysis. A total of 119,310 articles were considered so TM highly eased the arduous labor required if these articles were analyzed at hand. Scores were calculated according to the occurrence of words in the title, abstract, and indexing terms. The meta-analysis revealed that serious gastrointestinal events were very rare, but the use of aspirin was associated with a high risk of minor gastrointestinal in comparison with placebo or active comparators. Heintzelman et al. (2013) studied the importance of combining TM (Feldman & Sanger, 2007), NLP (Collobert et al., 2011) and UMLS Meta-thesaurus (Bodenreider, 2004) to properly classify and study different pains of patients with metastatic prostate cancer. According to the authors, a multiple regression model was constructed to assess the strength of associations between the occurrence of severe pain and all defined variables (e.g., receipt of various drugs). Results suggested the feasibility of tracking longitudinal patterns of pain by TM of free text clinical records. Patterns in the pain experience, undetectable without the use of NLP to mine the longitudinal clinical record, were consistent with clinical expectations. Cole et al. (2013) performed a retrospective cohort study with the support of TM techniques to analyze associations between allergic conditions and chronic uveitis in juvenile idiopathic arthritis patients. In order to extract patient characteristics from unstructured clinical notes different TM techniques (Gupta et al., 2009) were used, including identification of medical entities (e.g., diseases or drugs), detection of negations, standardization, and disambiguation of terms using 22 clinical ontologies. Results, based on a multivariate logistic regression model, reproduced four previously known associations, and presented a new association between allergic conditions and chronic uveitis in juvenile idiopathic arthritis patients. Leeper et al. (2013) performed a novel text-analytics pipeline to detect the adverse events associated with the use of Cilostazol to patients with a peripheral arterial disease. To analyze 1.8 million electronic records and select a cohort of patients with appropriate characteristics for this study, different TM tasks were performed, including drug or disease recognition, standardization, and detection of denied concepts. Authors carried out a multivariable logistic regression to determine the relationship between the treatment assignment (Cilostazol/No Cilostazol) and 18 covariates such as age, sex or hypertension, among others. Results

showed the nonassociation between the use of this drug and any major cardiovascular event (e.g., myocardial infarction, stroke, or death).

2.2.2 | Knowledge discovery

Tafti et al. (2017) presented the development of a Big Data Neural Network system whose main objective is the discovery and identification of adverse drug events from scientific articles and social networks related to health. To carry out this task, authors used a ML-based approach (Michalski et al., 2013), NLP (Collobert et al., 2011) and distributed processing frameworks such as Apache Spark. The proposal, named Word2Vector algorithm (Mikolov, Chen, Corrado, & Dean, 2013), is an algorithm based on Deep Learning (Shickel, Tighe, Bihorac, & Rashidi, 2017) that was evaluated obtaining a superior performance (Precision 0.936 and Recall 0.930) than traditional models (e.g., BOW [Zhang et al., 2010] + Decision tree [Sebastiani, 2002] with a Precision value of 0.875 and a Recall value of 0.872). The most interesting contribution of this system was the discovery of new rare side effects (e.g., lactic acidosis caused by metformin use). Byrd et al. (2014) presented a hybrid system, which is based on rule-based NLP and ML (Michalski et al., 2013), to detect signs and symptoms on clinical text of primary care that reveal the onset or development of a heart disease. The principal aim was to discover and identify 15 over 17 Framingham criteria (one of the most used in the diagnosis of heart failure), only based on clinical notes of primary care. Authors considered the Unstructured Information Management Architecture (UIMA) framework (Ferrucci & Lally, 2004) for the preprocessing of clinical reports and text analysis tasks. The overall performance of the proposed system slightly exceeded the standard (0.911 vs. 0.898 in *F*-measure). Collier (2012) provided an overview of the role played by TM techniques (Gupta et al., 2009) to discover novel information from large-scale text collections, in particular in epidemic detections. Author analyzed the Biocaster tool (Collier et al., 2008), a really interesting web service based on NLP (Collobert et al., 2011) and TM techniques (Gupta et al., 2009), which was able to improve the early detection of outbreaks of infectious diseases through linguistic signals detected on the web (e.g., forums, social networks, news). Authors used TM tasks like entity recognition, topic classification, and disease/location detection. They considered a Naive Bayes algorithm for automatic classification of the reports for topical relevance, achieving a high performance (a *F*-measure value of 0.930). Yang et al. (2017) presented an interesting and novel system that, just considering the admission notes, it was able to discover and infer the medication that the patient should take at the medical discharge. To carry out this task, authors used a deep learning (Shickel et al., 2017) approach based on a convolutional neural network model (Kim, 2014). The method was evaluated against four baseline systems (multilayer perceptron, SVM, random forest, and logistic regression) (Kotsiantis, 2007), achieving an improvement of 20% in the macro-averaged *F*-measure.

2.3 | Text summarization

Summarization aims at identifying the main topics of a document in order to make a summary that includes its key points. This is therefore a really important technique that allows the important information to be read by healthcare professionals in a reduced quantum of time, decreasing therefore the personnel costs and making it less sensible to the subjectivity that appears when large volumes of information are analyzed. Summarization techniques (Afantenos et al., 2005) can be grouped into different categories (single or multidocument summarization, text or multimedia summarization, general purpose or domain-specific, among others). Nevertheless, according to Hahn and Mani (2000) , the two main groups for these techniques are extractive and abstractive summarization.

2.3.1 | Extractive methods

These methods can be defined as the discovery of a collection of terms, phrases, or paragraphs that are highly representative of the context of the original text. Elhadad et al. (2005) presented an extractive and multidocument summarization system integrated in PERSIVAL (McKeown et al., 2001)(PErsonalized Retrieval and Summarization of Images, Video and Language), a framework that performs a different summary strategy depending on the type of user (physician or nonprofessionals). The expert user interacts with the system (arising questions in natural language) to look for patients with specific features. The search engine included in the system selects relevant multimedia documents, whereas a text summarizer module generates a multimedia summary with text, video, and images. The results demonstrated the effectiveness of the system (Precision 0.900 and Recall 0.650) in obtaining relevant information that support the decision-making process carried out by clinician. Sarkar et al. (2011) proposed an extractive summarization system that allows the automatic generation of summaries from medical news articles. To carry out this task, authors used an ML-based approach (Michalski et al., 2013). The system comprises three different phases: (a) document preprocessing; (b) a bagging meta-learner to extract sentences where the C4.5 algorithm (Quinlan, 1993) was considered as the base learner; and (c) a summary generation. The results showed that the proposed system performed better than the best of the baseline systems evaluated (Precision 0.590 and Recall 0.380 against Precision 0.540 and Recall 0.310).

2.3.2 | Abstractive methods

In abstractive summarization the synthesized information is presented as new text formed through the study and understanding of the semantics of the original text. Fiszman et al. (2009) proposed a methodology based on semantic abstraction to find relevant information about some specific diseases based on PubMed search results. Authors considered the SemRep (Rindflh & Fiszman, 2003), a NLP (Collobert et al., 2011) tool, to extract entities and relations from textual reports. Additionally, it makes use of the UMLS Metathesaurus (Bodenreider, 2004) and the MetaMap Transfer (Meystre & Haug, 2005) tool (https://mmtx.nlm.nih.gov/MMTx/) to recognize UMLS concepts. Authors compared their results with a baseline system, obtaining a better average Precision (a value of 0.390 vs. 0.170). Rindflesch et al. (2011) presented Semantic MEDLINE (https://skr3.nlm.nih.gov/SemMed/), a web application based on the SemRep system and the UMLS Metathesaurus that automatically summarizes all the MEDLINE citations returned by a PubMed search. Semantic MEDLINE provides four types of summaries: diagnosis; substance interaction; treatment of a disease; and pharmacogenomics. One of its major features is the resulting summarization, which is shown in the form of a graph (see Figure 3) to ease its comprehensibility. This figure graphically represents a summary of more than 780 MEDLINE articles dealing with Alzheimer's disease, showing in the nodes the clinical entities of interest (e.g., disorder or drugs) and the semantic relationships between these concepts (e.g., coexists with, location of, causes), discovering connections that in a manual way would go unnoticed. Zhang et al. (2011) proposed an automatic abstractive summarization system to improve the Semantic MEDLINE tool. According to the authors, the graphical summarization was not appropriate for more than 500 citations so it hampered future research studies. To solve this issue, authors considered the degree centrality method (Erkan & Radev, 2004) to reduce the number of nodes by measuring the number of edges connected to each node. This methodology was evaluated on 50,000 citations related to different diseases (Alzheimer, migraine, peptic ulcer, heart failure, melanoma, among others). The results of the proposed system were compared to those obtained by the reference standard (produced by physicians), demonstrating that the overall performance of the proposal was significantly better than the baseline system (0.720 vs. 0.470 in *F*-measure).

2.4 | Terminology extraction

Ontologies, thesauruses, corpus, and specialized databases are very important resources in medicine since they provide terminological and conceptual references to the application domains, enabling different tasks related to the natural language to be automatized. The development of such important linguistic resources is an arduous task that requires different techniques and

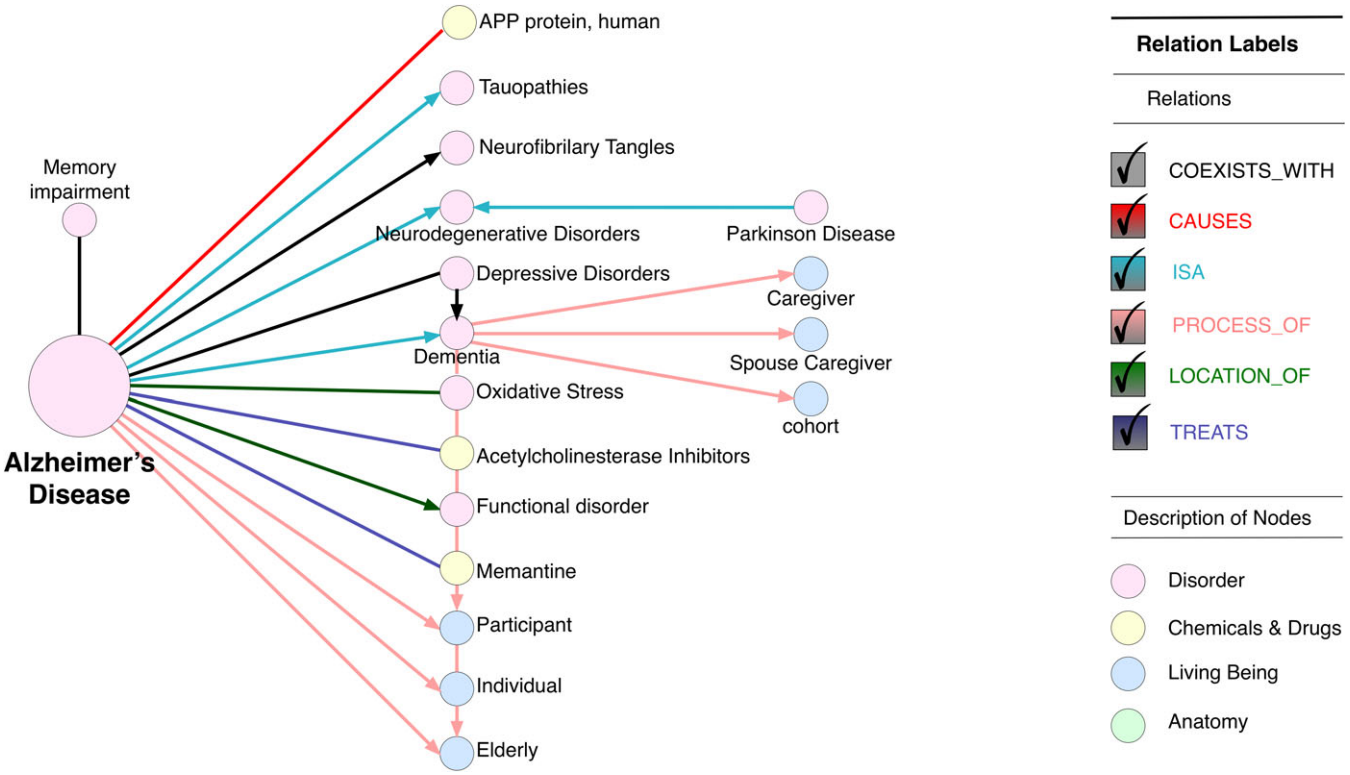

FIGURE 3 Semantic MEDLINE graph. Example of summarization according to treatment of a disease summary type (query: Alzheimer's disease). The nodes represent different entities found in citations and the arrows the connections between those entities

tools to be used. Some research studies on this matter are analyzed and they denote how TM (Gupta et al., 2009) and NLP techniques (Collobert et al., 2011) may automatically produce corpus, ontologies, and specialized databases.

2.4.1 | Corpus construction

Roberts et al. (2009) described the construction of a semantically annotated corpus of clinical texts, known as CLEF (Roberts et al., 2007) corpus (http://nlp.shef.ac.uk/clef/). The annotation methodology is based on NLP (Collobert et al., 2011). The corpus consisted on structured records and unstructured documents related to 20,234 different patients who suffer from cancer. The unstructured documents included clinical narratives as well as histopathology and imaging reports. The main clinical entities and relationships were identified. For the sake of implementing the system, the GATE NLP toolkit (Cunningham, Maynard, Bontcheva, & Tablan, 2002) and UMLS (Bodenreider, 2004) were used. According to its authors, this corpus provided great benefits for the development of effective information extraction system.

Currently, most of existing corpus include annotations for a single type of entity and few annotate relationships between entities. The existence of systems that automatically extract these relationships from literature and scientific databases is of great utility. Van Mulligen et al. (2012) described a corpus, called EU-ADR (http://biosemantics.org/index.php/resources/euadr-corpus), containing annotations of multiple entities and relations. The authors focused on a set of entities (diseases, drug and targets) and interrelationships between them (target–disease, target–drug, and drug–disease). Authors used an automatic NER system for entity annotation that was based on a thesaurus. According to the authors, the construction of this corpus is crucial to train and evaluate TM systems.

2.4.2 | Ontologies construction

Fabian et al. (2012) developed a method to automatically extend ontologies. The main novelty lay in the inclusion of new terms taken from queries, and textual information taken from different sources (journal articles, patents, text books, wiki pages, etc.). Authors combined two approaches, one was based on the structure of HTML documents, and the other one was based on multiple TM techniques (Gupta et al., 2009). The results showed that the idea of combining both approaches improved the results (a Recall value of 0.800 and a Precision value of 0.610 were obtained) when comparing with the approaches in isolation. Luther et al. (2011) employed a statistical TM approach to develop a clinical vocabulary for posttraumatic stress disorder (PSTD) in Veterans Health Administration from outpatient progress notes. Authors used SAS Text Miner (Abell, 2014), a tool that includes different functionalities including text parsing and extraction, automatic text cleaning, categorization, text clustering and predictive modeling of textual data, among others. As a result, a vocabulary formed by 226 unique PTSD-related terms was obtained. The authors compared their results with those generated from three different sources: focus group, review of SNOMED (Donnelly, 2006) terms, and review of practice guidelines of PTSD. The performance of the proposed system was analyzed against the rest of the systems evaluated to detect different categories of concepts (symptoms, treatments, etc.), obtaining the highest value in unique terms discovery (23.0% of the unique terms found), which is even higher than the one obtained by SNOMED (22.4% of the terms found).

2.4.3 | Databases construction

Fang et al. (2008)) analyzed the construction of a database, named TCMGeneDIT (http://tcm.lifescience.ntu.edu.tw/), that included information about relationships between the traditional Chinese medicine (TCM), genes, diseases, and TCM effects obtained from research bibliography. The database contained 13,167 genes and 3,360 disease entries that were obtained from 38,072 MEDLINE abstracts. To carry out this task, authors used a TM-based approach (Allahyari et al., 2017) combined with a rule-based approach. Authors showed that the construction of this database facilitated the analysis of different associations between genes, diseases, or proteins, obtaining high Precision results (Gene 0.928 and Gene-Disease 0.870). Xie et al. (2013) developed a microRNA cancer association database called miRCancer. For the development of this database, authors proposed an approach based on rules and different TM techniques (Gupta et al., 2009). Authors used an International Classification of Diseases for Oncology (ICD-O) for the cancer name recognition and regular expressions to identify miRNA names. miRCancer obtained 878 pairs of miRNAcancer associations on more than 26,000 articles from PubMed. Authors compared the performance of the proposed system with miR2Disease, a manually curated database on miRNAcancer. Both systems obtained the same value in Precision (the maximum value was obtained, i.e., 1.000) but the proposed system obtained a higher value in Recall (0.785 vs. 0.770). According to the authors, the use of TM techniques (Gupta et al., 2009) enabled a minimization of the manual maintenance of the database.

2.5 | Text classification

Nowadays, automatic text classification has become an essential task in medicine especially due to the quantity of textual information available in many disparate sources (databases, articles, social networks, forums, news, etc.). In the medical

research literature, many applications of text classification can be found including clinical alerts or risk factors categorization (Pakhomov et al., 2007), adverse events classification (Harpaz et al., 2014), electronic health records classification (Castro et al., 2015), symptomatology categorization (Vijayakrishnan et al., 2014), health miner (opinion and sentiment analysis mining) (Asghar et al., 2013). Nevertheless, most of research studies are focusing on three main applications: automatic diagnostic classification; patient stratification; and classification of medical literature.

2.5.1 | Automatic diagnostic classification

Metais et al. (2006) presented a TM system, called CIREA project, with the purpose of automatizing ICD-10 (International Classification of Diseases, Tenth Revision) coding by considering both TM (Gupta et al., 2009) and ML (Michalski et al., 2013) techniques. Its main objective was to infer a diagnostic code ICD-10 based on the textual content of medical reports. To carry out this task authors used a rule-based approach considering ML. It was defined within the scope of multilabel classification (each clinical report can be identified with multiple diagnostic codes) by including a novel multilabel classification algorithm called CLO3, based on relationship between usage of terms and diagnostics. The system includes a preprocessing phase of the medical reports, where several standardization tasks are carried out, for example, stemming by using an adaptation of the Porter (1980) algorithm. For the evaluation of the CLO3 algorithm, authors faced their proposal against the Naive Bayes algorithm, obtaining as a result an improvement of 6.7% in *F*-measure for the classification of medical report. Goldstein et al. (2007) presented a system capable of automatically inferring ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes of radiology reports. To accomplish this task, authors proposed three different approaches: (a) a first one based on Lucene (https://lucene.apache.org/core/) (Bialecki, Muir, Ingersoll, & Imagination, 2012) (an open-source library that allows analysis of textual collections); (b) another based on BoosTexter (Schapire & Singer, 2000) (a boosting algorithm for text classification); and (c) a rules-based approach to detect negations, synonyms and other semantic components. Results showed that the three analyzed approaches significantly improved the predictive performance with respect to the baseline system. In the best case, a *F*-measure value of 0.886 was obtained (the baseline obtained a value of 0.241 in *F*-measure).

2.5.2 | Patient stratification

Zuccon et al. (2013) analyzed the task of classifying certain clinical characteristics of patients (fractures or other abnormalities) from free-text radiological reports. Authors used a varied set of ML algorithms (Sebastiani, 2002), Naive Bayes, SMO and SPegasos, a variation of SVM, to carry out this task. They considered some external tools, for example, SNOMED CT Thesaurus (Donnelly, 2006), for the extraction of concepts. Authors evaluated four different types of configurations (bigram, stem, etc) with different classification algorithms. The best performance was achieved by the configuration based on bigrams with the Naive Bayes algorithm (*F*-measure value of 0.932). In Jonnagaddala et al. (2015), authors presented a multiclass classification system to automatically identify smoking status (current smoker, past smoker, past or current smoker, nonsmoker, and unknown status) from unstructured electronic health records. The proposal included different NLP techniques (Collobert et al., 2011) such as stop words removal, tokenization, and stemming. Authors used a hybrid approach using rule-based and ML techniques (Michalski et al., 2013). Results revealed that the selection of features using topic models increased the performance of the classification (topic models, *F*-measure of 0.837; traditional feature, *F*-measure of 0.827; and baseline, *F*-measure of 0.819).

2.5.3 | Medical literature classification

Frunza et al. (2011) proposed a methodology to create an automated system to assist humans in the preparation of systematic reviews. The task of collecting thousands of articles and manually labeling them is an arduous process that consumes a lot of time and resources. For the sake of automating and easing the task, authors proposed a ML-based approach (Michalski et al., 2013) and three types of text representations: BOW (Zhang et al., 2010), UMLS concepts, and a combination of both. The data set used in the experimental analysis consisted of 47,274 abstracts obtained from MEDLINE, and the results demonstrated that CNB achieved really promising results (the highest obtained Recall value was 0.678, whereas the highest obtained Precision value was 0.379). Yetisgen-Yildiz and Pratt (2005) presented a text classification system that allowed the task of classifying medical literature from the MEDLINE database to be automated. Authors proposed the use of NLP techniques (Collobert et al., 2011) and an approach based on SVM (Yu et al., 2010) to categorize more than 180,000 MEDLINE documents. In order to increase the performance of the proposed task, authors proved different types of text representation such as BOW (Zhang et al., 2010), bag-of-phrases (El-Kishky, Song, Wang, Voss, & Han, 2014) and a hybrid method, formed by the combination of both. The results demonstrated that the proposed hybrid textual representation, combining the features extracted from the BOW and bag-of-phrases representation, offered a better performance in *F*-measure than the other approaches (hybrid approach, a value of 0.600; BOW approach, a value of 0.580; bag-of-phrases approach, a value of 0.570).

3 | TM RESOURCES FOR MEDICINE

All the existing TM techniques in medicine are required to be complemented with different resources and tools such as corpus, meta-thesauruses, ontologies, part-of-speech (POS) taggers, parsing tools, named entity and relation extractors, etc. It is therefore of high interest to describe such resources that are essential in any TM-based system (Feldman & Sanger, 2007), some of the most important ones are described in this section. Due to space limitation a more detailed description of multiple TM techniques in medicine has been included at http://www.uco.es/kdis/textminingmedicine/#resources.

3.1 | Terminological resources

According to literature and research studies, the use of corpora, ontologies and thesauri in medicine provide a series of advantages: standardization of the information; overcoming the language barrier; complex knowledge of a specific domain can be obtained; knowledge sharing and reusability; reduction of the terminological ambiguity; ability to be used in a varied set of heterogeneous systems.

GENIA (http://www.nactem.ac.uk/genia) is one of the most commonly used corpora on TM (Feldman & Sanger, 2007) in medicine. It is a corpus specifically developed to support the construction and evaluation of IE and TM (Feldman & Sanger, 2007) systems in the biomedical domain. *ONCOTERM* (http://www.ugr.es/~oncoterm/) is another important corpus designed as a complete repository of information about the complex terminology associated with cancer. Finally, *NCBI disease corpus* (https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/) is a relevant annotated corpus used to perform TM tasks, for example, disease NER. As for the ontologies used in the fiof TM and medicine, Disease Ontology (http://disease-ontology.org/) appears as really important one. It is an open source ontology consisting of 8,043 hereditary, developmental, and acquired human diseases. Another example of highly representative ontology is *GALEN* (http://bioportal.bioontology.org/ontologies/GALEN), an open ontology that includes anatomical concepts, diseases, symptoms, drugs, and procedures, as well as the existing relationships between entities. Finally, it is also important to highlight Ontology of Adverse Events (http://www.oae-ontology.org/), an ontology focused on the definition and classification of adverse events occurring after a medical act.

Regarding the most important thesauri in Medicine, UMLS (http://www.nlm.nih.gov/research/umls) is a repository of multiple controlled vocabularies (more than 150) in biomedical sciences and health care. MeSH (Medical Subject Headings) (http://www.ncbi.nlm.nih.gov/mesh) is a controlled vocabulary thesaurus for indexing and classifying biomedical and health-related information. SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms) (http://www.snomed.org/snomed$-$ct) is a multilingual clinical health care terminology that includes three types of component: concepts, descriptions, and relationships.

3.2 | Data preprocessing tools

This is a crucial task in which the set of textual documents is transformed into a set of structured information by means of the application of a series of techniques: stop word removal, tokenization, stemming, word tags, among others. The data preprocessing process can be automated thanks to different existing tools, GENIA sentence splitter (http://www.nactem.ac.uk/y-matsu/geniass/) being one of the most important ones since it carries out the process of segmenting the sentences as an input text. Another important tool is *GENIA tagger* (http://www.nactem.ac.uk/GENIA/tagger/), an English text tagger that also allows named entity recognition to be performed. *Stanford Part-of-Speech Tagger* (https://nlp.stanford.edu/software/tagger.shtml) is another important software that allows to read and segment a sentence into tokens as a part of speech tag (name, verb, adjective, etc.) can be assigned to each token. Stanford Parser (https://nlp.stanford.edu/software/lex-parser.shtml) is a statistical parser available for English, German, Chinese, and Arabic languages. Finally, *FreeLing* (http: //nlp.lsi.upc.edu/freeling/) is an open source library that allows to perform a wide range of functions for automatic multilanguage processing, including named entity detection, POS-tagging, parsing, disambiguation, among others.

3.3 | NER and relation extraction tools

Some of the most important tools for the extraction of named entities and the identification of relationships are described below. DNorm (https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/DNorm.html) is a really interesting tool capable of recognizing and normalizing named entities related to diseases. PolySearch (http://polysearch.cs.ualberta.ca/index) is a web server based on TM to discover associations between various types of biomedical entities (diseases, drug, genes, or adverse effects). MEDIE (http://www.nactem.ac.uk/medie/) is a semantic search engine capable of extracting biomedical correlations from more than 14 million articles in MEDLINE. BeCAS (Biomedical Concept Annotation System)

(http://bioinformatics.ua.pt/becas/) is an application program interface (API) whose main objective is the automatic identification and annotation of biomedical concepts (disorders, anatomical concepts, genes, biological processes).

3.4 | Advanced TM tools

Plenty of tools that integrate and unify some of the TM tasks (tokenization, stemming, detection and extraction of named entities, etc.) have been described in literature. One of these tools is MetaMap (https://metamap.nlm.nih.gov/), a widely used tool in medicine which main aim is to find relevant concepts in a wide collection of biomedical texts by using the UMLS metathesaurus as terminological and semantical basis (Bodenreider, 2004). MetaMap includes different NLP techniques (Collobert et al., 2011), and it is able to perform multiple tasks such as detection of negated terms, words disambiguation, or acronyms detection. UIMA (https://uima.apache.org/external-resources.html) is another notorious software tool whose main objective is to analyze unstructured information in order to discover relevant data for the end user. Among its principal functionalities the following can be highlight: named-entity detectors, analysis of dependencies, grammatical parsing, annotation, document classification, and multilingual analysis. Apache cTAKES (clinical Text Analysis and Knowledge Extraction System) (http://ctakes.apache.org/) is an open source system specifically designed for the extraction of relevant information from electronic medical records. It offers a great variety of functionalities for the analysis of texts and the extraction of information in medicine: negation detection, NER, drug mention annotation, coreference detection, etc.

4 | OPEN CHALLENGES IN CLINICAL MEDICINE

The continuous advance of TM and its applications to medicine has been a great support for both patients and health professionals. Some of the most important benefits achieved by the application of TM techniques (Gupta et al., 2009) in the clinical field can be summarized as follows: improvement in the quality of care services (Delespierre, Denormandie, Bar-Hen, & Josseran, 2017); reduction in the number of medical errors (Cohan, Fong, Ratwani, & Goharian, 2017); supporting the prevention and detection of diseases (Just, 2017); identification of the most beneficial treatments (Palanisamy & Thirunavukarasu, 2017); and improving both the time and costs related to the health management (Sun et al., 2017). A perfect equilibrium between offering top quality care services and getting it at the lowest possible cost is precisely the major current challenge of health institutions. Hence, it is necessary to incorporate novel technologies in the health institutions so a more personalized, predictive, and effective medicine can be approached. The present and future road map to achieve these objectives is mainly marked by TM techniques (Gupta et al., 2009) as described below.

4.1 | Advanced systems for clinical decision support

Clinical decision support systems are the result of the synergy of disciplines such as TM, ML (Michalski et al., 2013), and medicine. The development of these systems is not novel, however, the real installation and start-up of these systems in health centers does not go hand-in-hand with the progress in the research and theoretical developments of these tools. This is generally caused by the peculiarities of the domain, which includes, among others, technological backwardness of hospitals, lack of computer knowledge as a barrier for health professionals, and a great amount of information written in natural language with lacking standardization and structuring. All of this makes essential to provide mechanisms that enable the use of these tools in the daily clinical practice to be increased, proposing reliable, intuitive, and fast response systems.

Some existing solutions based on TM have been already proposed by different researchers. Watson Health (Ferrucci, Levas, Bagchi, Gondek, & Mueller, 2013) (https://www.ibm.com/watson/health/) is a clear example of how the synergy of various disciplines such as TM, NLP (Collobert et al., 2011) and ML (Michalski et al., 2013) can help in building cognitive systems that support the health care professionals in complex tasks such as diagnostic prediction, automatic creation of individualized treatment plans, infer individual health needs of a patient, etc. It was specifically created to be used on oncology, but it is now a reality in different health care institutions in helping physicians with the process of choosing the most appropriate treatment. Watson Health is based on textual information from diverse medical records (more than 15 million articles, 200 books, and 300 medical journals) to infer the knowledge needed to generate recommendations. Currently, there are some open research lines that are being addressed with Watson Health: evidence-based cancer care, drug discovery, clinical trial matching, and risk stratification.

Advanced systems for supporting the clinical decision is specially alluring in the emergency department. It is, perhaps, the one that requires the most accurate solution as fast as possible due to the situation is crucial. The use of TM has played an important role in the development of intelligent systems that support decision making in the emergency services, and its application is already an incipient reality. In Portela et al. (2013), authors described the development of a specific system for

emergency services that guides the healthcare professional in a correct decision-making process to establish clinical priorities . This complex process was carried out thanks to TM techniques (Gupta et al., 2009) that extract relevant data from electronic medical records, laboratory tests, or therapeutic plans. According to the authors and the tests performed in a hospital, the proposed system enabled an optimization of the resources and a reduction in the waiting time.

4.2 | Health social media

The analysis of textual information in social networks and virtual communities related to health is an important source of knowledge to increase effectiveness in healthcare. With the help of this valuable information it is possible to carry out public health surveillance tasks, to evaluate epidemiological risks, and even to detect health alerts. The challenge of conducting this in-depth analysis of social networks can be overcome with the use of TM techniques (Gupta et al., 2009). As an example of this applicability, let us consider the analysis of tweet contents, which provide interesting health behavioral profiles from a population. Yoon, Elhadad, and Bakken (2013) considered the physical activity as the main topic to be analyzed within tweet contents, denoting the healthy habits to help in the prevention of diseases, for example, cardiovascular problems. In a similar way, Paul and Dredze (2012) analyzed more than 1.5 million tweets and discovered mentions of mild, acute, and chronic illnesses (e.g., obesity, insomnia, and allergies). They also analyzed symptoms and medications, placing all these illnesses by geographical areas.

According to the authors, their research work may be of great support in syndromic surveillance and can serve as a guide in the generation of new hypotheses. Corley, Cook, Mikler, and Singh (2010)) provided a disease surveillance resource that was able to identify diseases in online communities. Authors used TM techniques (Gupta et al., 2009) and Spinn3r (http://docs.spinn3r.com/#overview), a web service for indexing social media, blogs, news, etc. Their aim was to determine outbreaks of influenza from information contained in blogs. They also considered the SUBDUE (Cook & Holder, 1994) algorithm, a graph-based data mining model to identify anomalies in flu from blogs.

5 | LESSON LEARNED

TM gathers the most useful techniques to derive high-quality structured information from unstructured textual data, which is specially relevant in medical data since clinical information tends to be ambiguous and includes complex vocabulary; does not follow a terminological standardization; includes sentences that may contain grammatical errors and acronyms; and it is usually written in a narrative form. In the performed review, the use of TM in medicine has been described as an important process on five different applications: NER; hypothesis generation and knowledge discovery; text summarization; terminology extraction and text classification. Through this review, different ways in which TM has contributed to further analysis in medicine are described.

Focusing on NER, it has been mainly used to extract medical entities from medical texts (e.g., electronic health records, medical literature, health social networks) and most of existing approaches aimed at finding concepts (e.g., diseases, drugs, symptoms), time expressions (e.g., date of onset of a disease, duration of a treatment) as well as semantic relationships (e.g., disease–treatment, disease–test, disease–symptom). NER has been mainly applied to English texts, some promising results have been also obtained on different languages such as Swedish and Spanish, and it has been really relevant in supporting the detection of medical problems found in discharge reports; the terminological standardization; the discovery of relationships between medical problems and treatments; as well as the anonymization of personal data in clinical documents, among others.

Another important application of TM in medicine is to provide valuable and useful information to support the discovery of new hypotheses and hidden knowledge on textual data. TM has been used as a base procedure to automatize and ease the process of grading hundreds of articles for relevance to a posterior meta-analysis that enabled to discover that the use of aspirin was associated with a high risk of minor gastrointestinal in comparison with placebo or active comparators. It has also been used to extract medical entities that were then considered in a logistic regression to properly classify different pain statuses of patients with metastatic prostate cancer; pain statuses with receipts of various drugs; and cardiovascular events (e.g., myocardial infarction, stroke, or death) with a specific drug. TM has also been considered as a previous process for applying ML algorithms on medical data. Among the findings in this regard, it is possible to list the detection of different signs and symptoms that revealed the onset or development of a heart disease; the discovery of outbreaks of infectious diseases through the web (e.g., forums, social networks, news); and the deduction of the specific medication that a patient should take at the medical discharge, among others.

Text summarization is another application of TM in medicine through which the main topics of a document are identified to make a summary. It therefore eases the labor of the health care professionals that need to read tons of research articles,

reducing the time, decreasing the costs, and being less sensitive to subjectivity. Two methods of summarizing clinical texts (extractive and abstractive) have been widely used in medicine, supporting the construction of a digital library based on the summarization of relevant medical literature; being useful for a novel system that summarizes scientific articles as well as for different visual tools that synthesizes different clinical concepts and their interrelations.

As for terminology extraction, the application of TM techniques in medicine is enabling an increase in both number of resources and quality of such resources. This increment has resulted in a much easier development and evaluation of TM-based systems. Research studies in this regard have contributed to the construction of clinical vocabularies and the development of specialized databases in the biomedical domain. Finally, text classification is an essential task in medicine especially due to the large amounts of textual information available in many disparate sources (databases, articles, social networks, forums, news, etc.). Text classification has been useful to automatize the daunting process of assigning diagnostic codes to medical records; to identify groups of patients based on their life habits or health problems; as well as to categorize scientific articles and medical literature in order to ease the clinical research work.

With the help of this review, it has been demonstrated the continuous advance of TM in medicine has constituted a great support for both patients and health professionals in terms of improvement in the quality of care services; reduction in the number of medical errors; prevention and detection of diseases; identification of the most beneficial treatments; reduction of time and costs related to the health management. Finally, some tips about future challenges have also been given, including advanced systems for clinical decision support as well as the analysis of textual information in social networks and virtual communities.

6 | CONCLUSIONS

This paper has provided an extensive review about TM techniques, performing an analysis, and description of more than 90 research papers, with special attention on those related to the applicability of TM in different areas of medicine. In this regard, many practical applications have been described, presenting the methods, techniques, tools, and obtained results for each of the analyzed research studies. All these findings have been summarized into a summary table that is available at http://www.uco.es/kdis/textminingmedicine/#summarytable due to space limitations. In this review, therefore, the impact of TM techniques in medicine has been highlighted, improving the early disease diagnosis, developing novel and improved therapies that reduce risks and derived problems, producing new medical hypothesis, etc. Additionally, a set of TM resources for medicine have been denoted so that they can complement the techniques previously analyzed. Finally, different open challenges in clinical medicine have been described and analyzed.

ACKNOWLEDGMENTS

This work was Supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund, under the project TIN2017-83445-P.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

FURTHER READING

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. *Briefing in Bioinformatics 6* (3), 239-251. https://doi.org/10.1093/bib/6.3.239

REFERENCES

Abell, M. (2014). *SAS text miner*. Scotts Valley, CA: CreateSpace Independent Publishing Platform.

Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical doc-uments: A survey. *Artificial Intelligence in Medicine*, *33*(2), 157–177.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A brief survey of text mining: Classification, clustering and extraction techniques.* arXiv preprint arXiv:170702919.

Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229–236.

Asghar, M. Z., Qasim, M., Ahmad, B., Ahmad, S., Khan, A., & Khan, I. A. (2013). Health miner: Opinion extraction from user generated health reviews. *International Journal of Academic Research*, *5*(6), 279–284.

Baron, J. A., Senn, S., Voelker, M., Lanas, A., Laurora, I., Thielemann, W., … McCarthy, D. (2013). Gastrointestinal adverse effects of short-term aspirin use: A meta-analysis of published randomized controlled trials. *Drugs in R&D*, *13*(1), 9–16.

Benton, A., Hill, S., Ungar, L., Chung, A., Leonard, C., Freeman, C., & Holmes, J. H. (2011). A system for de-identifying medical message board text. *BMC Bioinformatics*, *12*(3), 1–10. https://doi.org/10.1186/1471-2105-12-S3-S2

Bialecki, A., Muir, R., Ingersoll, G., & Imagination, L. (2012). *Apache lucene 4.* Paper presented at SIGIR 2012 Workshop on Open Source Information Retrieval (p. 17), Portland, OR.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl 1), D267–D270.

Byrd, R. J., Steinhubl, S. R., Sun, J., Ebadollahi, S., & Stewart, W. F. (2014). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International Journal of Medical Informatics*, *83*(12), 983–992.

Carrero, F., Cortizo, J. C., & Gomez, J. M. (2008). *Building a Spanish MMTx by using automatic translation and biomedical ontologies.* Paper presented at International Conference on Intelligent Data Engineering and Automated Learning (pp 346–353), Springer, Berlin, Germany.

Castro, V. M., Minnier, J., Murphy, S. N., Kohane, I., Churchill, S. E., Gainer, V., … Belliveau, R. A., Jr. (2015). Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry*, *172*(4), 363–372.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated fi and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310.

Chen, H., Fuller, S. S., Friedman, C., & Hersh, W. (2005). *Knowledge management, data mining, and text mining in medical informatics.* Paper presented at Medical Informatics (pp 3–33), Springer, Bostan, MA.

Cohan, A., Fong, A., Ratwani, R. M., & Goharian, N. (2017). Identifying harm events in clinical care through medical narratives. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 52–59). Boston, MA: ACM.

Cole, T. S., Frankovich, J., Iyer, S., LePendu, P., Bauer-Mehren, A., & Shah, N. H. (2013). Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: A new model for EHR-based research. *Pediatric Rheumatology*, *11*(1), 45.

Collier, N. (2012). Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health*, *7*(7), 731–749.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., … Taniguchi, K. (2008). BioCaster: Detecting public health rumors with a web-based text mining system. *Bioinformatics*, *24*(24), 2940–2941.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*(August), 2493–2537.

Cook, D. J., & Holder, L. B. (1994). Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, *1*, 231–255.

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of infl za mentions in web and social media. *International Journal of Environmental Research and Public Health*, *7*(2), 596–615.

Cunningham, H., Maynard, D., Bontcheva, K, & Tablan, V. (2002). *A framework and graphical development environment for robust NLP tools and applications.* Paper presented at ACL (pp. 168–175)

Dandapat, S., & Way, A. (2016). Improved named entity recognition using machine translation-based cross-lingual information. *Computacion y Sistemas*, *20*(3), 495–504.

Delespierre, T., Denormandie, P., Bar-Hen, A., & Josseran, L. (2017). Empirical advances with text mining of electronic health records. *BMC Medical Informatics and Decision Making*, *17*(1), 127.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*, *121*, 279.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5), 352–359.

Elhadad, N., Kan, M. Y., Klavans, J. L., & McKeown, K. (2005). Customization in a unifi framework for summarizing medical literature. *Artificial Intelligence in Medicine*, *33*(2), 179–198.

El-Kishky, A., Song, Y., Wang, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, *8*(3), 305–316.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Fabian, G., W¨achter, T., & Schroeder, M. (2012). Extending ontologies by fi siblings using set expansion techniques. *Bioinformatics*, *28*(12), i292–i300.

Fang, Y. C., Huang, H. C., Chen, H. H., & Juan, H. F. (2008). TCMGeneDIT: A database for associated traditional Chinese medicine, gene and disease information using text mining. *BMC Complementary and Alternative Medicine*, *8*(1), 58.

Feldman, K., Hazekamp, N, & Chawla, N. V. (2016). Mining the clinical narrative: All text are not equal. Paper presented at 2016 I.E. International Conference on Healthcare Informatics (ICHI) (pp. 271–280), IEEE

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York, NY: Cambridge University Press.

Ferrandez, O., South, B. R., Shen, S., & Meystre, S. M. (2012). A hybrid stepwise approach for deidentifying person names in clinical documents. In *Proceedings of the 2012 workshop on biomedical natural language processing* (pp. 65–72). Stroudsburg, PA: Association for Computational Linguistics.

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3–4), 327–348.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: Beyond jeopardy! *Artificial Intelligence*, *199*, 93–105.

Fiszman, M., Demner-Fushman, D., Kilicoglu, H., & Rindflesch, T. C. (2009). Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of Biomedical Informatics*, *42*(5), 801–813.

Frunza, O., Inkpen, D., Matwin, S., Klement, W., & O'blenis, P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, *51*(1), 17–25.

Goldstein, I., Arzumtsyan, A., & Uzuner, O. (2007). *Three approaches to automatic assignment of ICD-9-CM codes to radiology reports.* Paper presented at AMIA Annual Symposium Proceedings, American Medical Informatics Association (p 279), vol. 2007.

Gupta, V., & Lehal, G. S., (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. *Computer*, *33*(11), 29–36.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., … Shah, N. H. (2014). Text mining for adverse drug events: The promise, challenges, and state of the art. *Drug Safety*, *37*(10), 777–790.

Heintzelman, N. H., Taylor, R. J., Simonsen, L., Lustig, R., Anderko, D., Haythornthwaite, J. A., … Bova, G. S. (2013). Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *Journal of the American Medical Informatics Association*, *20*(5), 898–905.

Huang, C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Briefings in Bioinformatics*, *17*(1), 132–144.

Jonnagaddala, J., Dai, H. J., Ray, P., & Liaw, S. T. (2015). A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. *ACL-IJCNLP*, *2015*, 147–151.

Just, E. (2017). *How to use text analytics in healthcare to improve outcomes—Why you need more than NLP. Health catalyst data: quality, management, governance.* Retrieved from: https://wwwhealthcatalystcom/how-to-use-text-analytics-in-healthcare-to-improve-outcomes

Kim, Y. (2014). *Convolutional neural networks for sentence classification.* arXiv preprint arXiv:14085882

Kipper-Schuler, K., Kaggal, V., Masanz, J., Ogren, P., & Savova, G. (2008). *System evaluation on a named entity corpus from clinical notes.* Paper presented at: Language Resources and Evaluation Conference, LREC (pp. 3001–3007)

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conference on Machine Learning* (pp. 282–289). San Francisco, CA: Morgan Kaufmann.

Leeper, N. J., Bauer-Mehren, A., Iyer, S. V., LePendu, P., Olson, C., & Shah, N. H. (2013). Practice-based evidence: Profiling the safety of cilostazol by text-mining of clinical notes. *PLoS One*, *8*(5), e63499.

Lin, Y. K., Chen, H., & Brown, R. A. (2013). MedTime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics*, *46*, S20–S28.

Luther, S., Berndt, D., Finch, D., Richardson, M., Hickling, E., & Hickam, D. (2011). Using statistical text mining to supplement the development of an ontology. *Journal of Biomedical Informatics*, *44*, S86–S93.

McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit*. Massachusetts: Mallet.

McKeown, K. R., Chang, S. F., Cimino, J., Feiner, S., Friedman, C., Gravano, L., et al. (2001). PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 331–340). New York, NY: ACM.

Metais, E., Nakache, D., & Timsit, J. F. (2006). Automatic classification of medical reports, the cirea project. In *Proceedings of the 5th WSEAS international conference on telecommunications and informatics* (pp. 354–359). Istanbul, Turkey: World Scientific and Engineering Academy and Society (WSEAS).

Meystre, S., & Haug, P. J. (2005). Evaluation of medical problem extraction from electronic clinical documents using MetaMap transfer (MMTx). *Studies in Health Technology and Informatics*, *116*, 823–828.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Berlin Heidelberg: Springer Science & Business Media.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:13013781

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Pakhomov, S., Weston, S. A., Jacobsen, S. J., Chute, C. G., Meverden, R., Roger, V. L., et al. (2007). Electronic medical records for clinical research: Application to the identification of heart failure. *The American Journal of Managed Care*, *13*(6 Part 1), 281–288.

Palanisamy, V., & Thirunavukarasu, R. (2017). Implications of big data analytics in developing healthcare frameworks–A review. *Journal of King Saud University-Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2017.12.007

Paul, M. J., & Dredze, M. (2012). A model for mining public health topics from twitter. *Health*, *11*, 16–16.

Portela, F., Cabral, A., Abelha, A., Salazar, M., Quintas, C., Machado, J., … & Santos, M. F. (2013). Knowledge acquisition process for intelligent decision support in critical health care. In *Information Systems and Technologies for Enhancing Health and Social Care* (pp. 55–68). Hershey, PA: IGI Global.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In IOS Press (Ed.), *Natural language processing using very large corpora* (pp. 157–176). Dordrecht: Springer.

Rindflesch, T. C., Kilicoglu, H., Fiszman, M., Rosemblat, G., & Shin, D. (2011). Semantic MED-LINE: An advanced information management application for biomedicine. *Information Services & Use*, *31*(1–2), 15–21.

Rindflh, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, *36*(6), 462–477.

Rink, B., Harabagiu, S., & Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, *18*(5), 594–600.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., et al. (2007). The CLEF corpus: Semantic annotation of clinical text. In American Medical Informatics Association (Ed.), *AMIA annual symposium proceedings* (Vol. 2007, p. 625). Bethesda, MA: American Medical Informatics Association.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, *42*(5), 950–966.

Roberts, K., & Harabagiu, S. M. (2011). A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, *18*(5), 568–573.

Roberts, K., Rink, B., & Harabagiu, S. M. (2013). A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association*, *20*(5), 867–875.

Sarkar, K., Nasipuri, M., & Ghose, S. (2011). Using machine learning for medical document summarization. *International Journal of Database Theory and Application*, *4*(1), 31–48.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507–513.

Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning*, *39*(2–3), 135–168.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604.

Skeppstedt, M., Kvist, M., Nilsson, G. H., & Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, *49*, 148–158.

Sun, W., Cai, Z., Liu, F., Fang, S., & Wang, G. (2017), A survey of data mining technology on electronic medical records. Paper presented at 2017 I.E. 19th International Conference on e-Health Networking, Applications and services (Healthcom) (pp. 1–6). https://doi.org/10.1109/HealthCom. 2017.8210774

Tafti, A. P., Badger, J., LaRose, E., Shirzadi, E., Mahnke, A., Mayer, J., … Peissig, P. (2017). Adverse drug event discovery using biomedical literature: A big data neural network adventure. *JMIR Medical Informatics*, *5*(4), e51.

Uzuner, O., Mailoa, J., Ryan, R., & Sibanda, T. (2010). Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*, *50*(2), 63–73.

Van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., … Furlong, L. I. (2012). The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, *45*(5), 879–884.

Vijayakrishnan, R., Steinhubl, S. R., Ng, K., Sun, J., Byrd, R. J., Daar, Z., et al. (2014). Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of Cardiac Failure*, *20*(7), 459–464.

Wang, Y., & Patrick, J. (2009). *Cascading classifiers for named entity recognition in clinical notes.* Paper presented at Proceedings of the Workshop on Biomedical Information Extraction, Association for computational linguistics (pp. 42–49)

Xia, Y., Zhong, X., Liu, P., Tan, C., Na, S., Hu, Q., & Huang, Y. (2013). Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1. Paper presented at: CLEF (Working Notes).

Xie, B., Ding, Q., Han, H., & Wu, D. (2013). miRCancer: A microRNA–cancer association database constructed by text mining on literature. *Bioinformatics*, *29*(5), 638–644.

Yang, Y., Xie, P., Gao, X., Cheng, C., Li, C., Zhang, H., & Xing, E. (2017). *Predicting discharge medications at admission time based on deep learning.* arXiv preprint arXiv:171101386

Yetisgen-Yildiz, M., & Pratt, W. (2005). *The effect of feature representation on MEDLINE document classification.* Paper presented at AMIA Annual Symposium Proceedings, American Medical Informatics Association (p. 849), vol. 2005

Yoon, S., Elhadad, N., & Bakken, S. (2013). A practical approach for content mining of tweets. *American Journal of Preventive Medicine*, *45*(1), 122–129.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and prediabetes. *BMC Medical Informatics and Decision Making*, *10*(1), 16.

Zhang, H., Fiszman, M., Shin, D., Miller, C. M., Rosemblat, G., & Rindflesch, T. C. (2011). Degree centrality for semantic abstraction summarization of therapeutic studies. *Journal of Biomedical Informatics*, *44*(5), 830–838.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52.

Zhu, X., Cherry, C., Kiritchenko, S., Martin, J., & De Bruijn, B. (2013). Detecting concept relations in clinical text: Insights from a state-of-the-art model. *Journal of Biomedical Informatics*, *46*(2), 275–285.

Zuccon, G., Wagholikar, A. S., Nguyen, A. N., Butt, L., Chu, K., Martin, S., & Greenslade, J. (2013). *Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the SNOMED CT ontology.* Paper presented at: AMIA Summits on Translational Science Proceedings 2013 (p. 300).

How to cite this article: Luque C, Luna JM, Luque M, Ventura S. An advanced review on text mining in medicine. *WIREs Data Mining Knowl Discov*. 2019;e1302. https://doi.org/10.1002/widm.1302