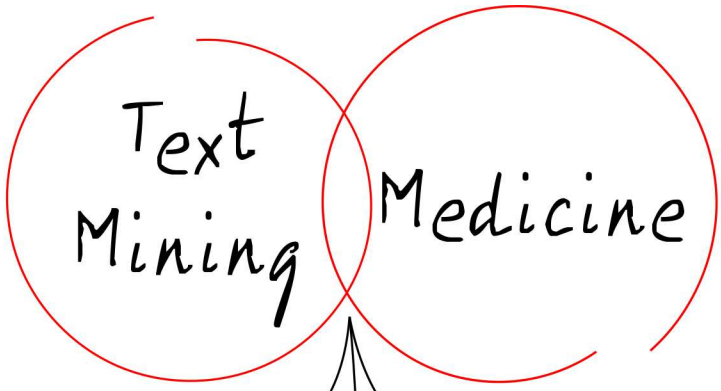
WIREs
DATA MINING AND KNOWLEDGE DISCOVERY
WILEY

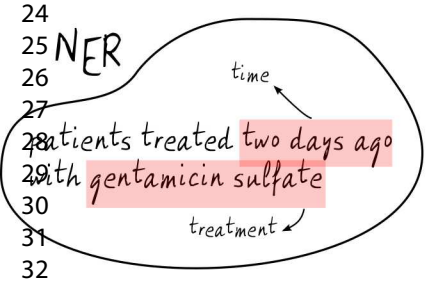
An advanced review on Text Mining in Medicine

Journal:	*WIREs Data Mining and Knowledge Discovery*
Manuscript ID	DMKD-00339.R1
Wiley - Manuscript type:	Advanced Review
Date Submitted by the Author:	n/a
Complete List of Authors:	Luque, Carmen; University of Cordoba, Computer Sciences and Numerical Analysis Luna, José; Universidad de Cordoba, Computer Science and Numerical Analysis Luque, María; University of Cordoba, Computer Sciences and Numerical Analysis Ventura, Sebastian; University of Cordoba, Computer Sciences and Numerical Analysis
Keywords:	
Choose 1-3 topics to categorize your article:	Biological Data Mining (DAAC) < Algorithmic Development (DAAA), Hierarchies and Trees (DAAD) < Algorithmic Development (DAAA), Model Combining (DAAE) < Algorithmic Development (DAAA)

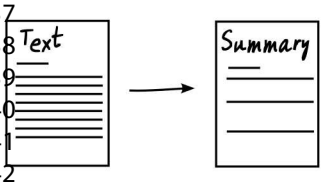


Applications

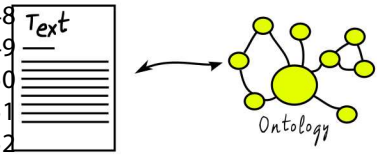
NER



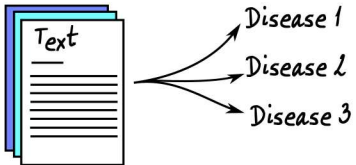
Text summarization



Terminology extraction



Text classification



Resources

GENIA

Unified Medical Language System

MetaMAP®

Unstructured Information Management Architecture
An Apache Project

cTAKES™

Open challenges

Advanced systems

IBM Watson

Health social Media

An advanced review on Text Mining in Medicine

C. Luque*, J. M. Luna†‡, M. Luque§, S. Ventura¶‖**

Abstract

Healthcare professionals produce abundant textual information in their daily clinical practice and this information is stored in many disparate sources and, generally, in textual form. The extraction of insights from all the gathered information, mainly unstructured and lacking in normalization, is one of the major challenges in computational medicine. In this respect text mining 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 revision about existing techniques and resources to perform text mining tasks in Medicine. In this review, more than ninety relevant research studies have been analysed, describing the most important practical applications, terminological resources, tools, and open challenges of text mining in Medicine.

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 healthcare professionals in their general daily practices[32]. The health of the patients is regularly described by thousands of doctors 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.

*Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory, Maimonides Biomedical Research Institute of Cordoba, Spain. Email: clgcordoba@gmail.com

†Knowledge Discovery and Intelligent Systems in Biomedicine Laboratory, Maimonides Biomedical Research Institute of Cordoba, Spain. Email: jmluna@uco.es

‡Department of Computer Science and Numerical Analysis, University of Cordoba, 14071 Cordoba, Spain. Email: jmluna@uco.es

§Department of Computer Science and Numerical Analysis, University of Cordoba, 14071 Cordoba, Spain. Email: mluque@uco.es

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

‖Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia. Email: sventura@uco.es

**Department of Computer Science and Numerical Analysis, University of Cordoba, 14071 Cordoba, Spain. Email: sventura@uco.es

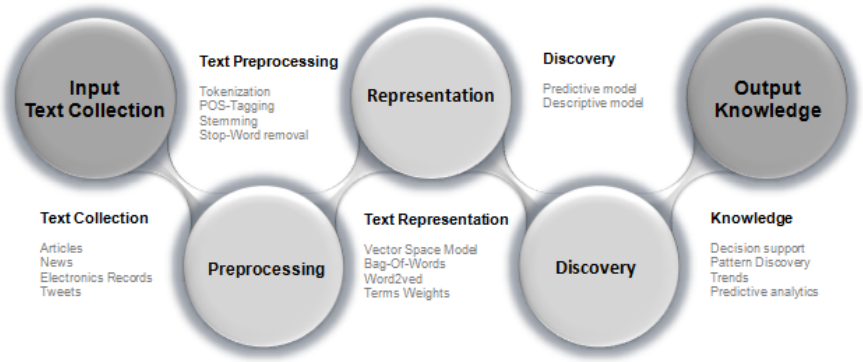


Figure 1: Text Mining Process.

Currently, the main problem to be faced by any healthcare 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 with 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 on 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 most useful techniques to derive high-quality structured information from unstructured textual data are gathered under the term Text Mining (TM)[33]. It is a process in which useful and unknown knowledge is extracted from textual data by applying a series of phases (see Figure 1)[33]: i) Preprocessing stage, where textual and unstructured input data are standardized and cleaned by means of different NLP techniques[19], e.g. tokenization and stemming; ii) 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, e.g. Bag-Of-Words (BOW)[93]; iii) Discovery stage, where useful, unexpected and unknown information is extracted from textual data collections through the application of certain methods and techniques[3], e.g. classification, clustering, etc. Due to space limitations, 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 healthcare professionals in both research and clinical daily issues, e.g. 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[40] have been widely studied and analysed from a technical perspective. Nevertheless, due to the increasing applicability of TM[33] to Medicine, it becomes essential to provide a general analysis of existing approaches and methodologies that have contributed to improve the healthcare. In this regard, the aim of this paper is to review the most interesting

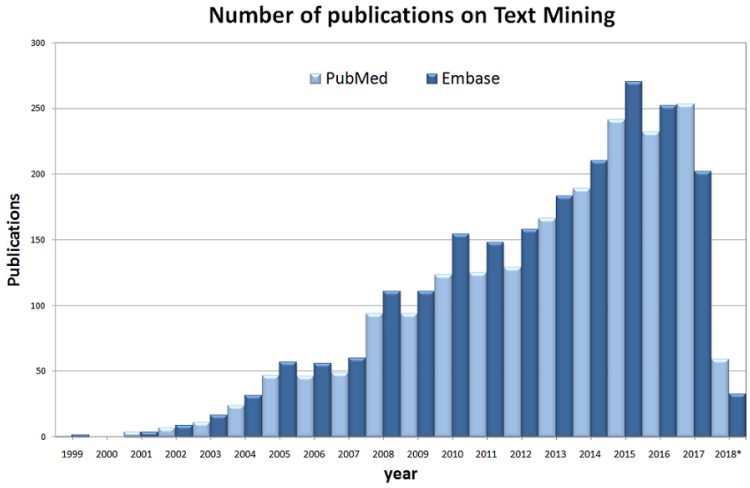


Figure 2: Number of scientific publications returned by PubMed and Embase biomedical databases (key words: Text Mining. Last updated March 2018).

research works published on the use of TM[33] in Medicine. More than ninety research articles have been analysed with special attention to those involving important applications such as named entity recognition, 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 rest of the paper is structured as follows. Section 2 describe the main contributions of TM in Medicine. Some terminological resources and tools used in textual analysis are also described in Section 3. Finally, Section 4 summarizes some open challenges, and Section 5 presents some concluding remarks.

Text Mining Applied to Medicine

Since the beginning of the century, when first research works appeared, TM techniques[40] applied to healthcare tasks (diagnosis, treatment and prevention of different diseases) are denoting an increasing interest[14], as it is demonstrated by the growth number of articles published in this regard on Embase and PubMed databases (see Figure 2). 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 limitations, the summary table is available at `http://www.uco.es/kdis/textminingmedicine/#summarytable`.

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. Named entity recognition (NER) is therefore the process of discovering named entities in different chunks

Table 1: Summary of references where Text Mining was applied to Medicine.

Applications	References
Named entity recognition	7,11,34,47,51,69,72,73,79,82,85,86,94
Hypothesis generation and knowledge discovery	6,10,16,17,43,50,81,88
Text summarization	2,28,37,68,74,92
Terminology extraction	30,31,52,71,83,87
Text classification	5,12,38,39,42,44,55,60,84,89,95

of textual data[59]. In the medical field, NER systems[85] 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[80]: rule-based, dictionary-based, Machine Learning(ML)-based[57], and a hybridization of the aforementioned approaches. To demonstrate the importance of NER in Medicine, a wide number of research works are analysed. First, the focus is on the extraction of medical concepts from clinical reports written in different languages. Then, the discovery of concepts related to temporal expressions is analysed. Later, the attention is paid on personal data anonimization. Finally, the extraction of relationships between entities is analysed.

Medical concepts from clinical reports. *Roberts and Harabagiu*[72] proposed the extraction of medical concepts (e.g. disease, drug, injury) in clinical texts under a ML-based approach[57]. Authors presented a classification task by considering two ML classifiers: Support Vector Machine (SVM)[91] and Conditional Random Fields (CRF)[49]. Additionally, the NegEx[13] algorithm was used to detect negations. For concept extraction, several external sources of knowledge were used including MetaMap[4], 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*[47] analysed 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[9], which was also expanded with synonyms. Authors compared the dictionary look-up model to different ML algorithms[57] (CRF[49] and SVM[91]). 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*[86] addressed the task of disease recognition from clinical texts that was proposed in the CLEF eHealth 2013 conference. Authors used a dictionary-based approach, combining MetaMap[4] and cTAKES[75] 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[23], some authors have started to explore their application on different languages. *Skeppstedt et al*[79] analysed 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 (IOB) encoding[66] for annotated entities and the CRF[49] algorithm, obtaining similar results (a F-measure value of 0.810 for disorder recognition) to those obtained on English texts. *Carrero et al*[11] 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)[56] 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%).

Time expressions. The accurate extraction of time expressions in the medical field (date of onset of a disease, duration and frecuency of the treatment, among others) is a really important and arduous task that have been widely studied. *Lin et al*[51] presented the MedTime system, a temporal information extraction system including different rule-based and ML[57] procedures (SVM[91] and CRF[49]). 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[4] for extracting features of medical lexicon and semantic types, and Mallet[53] 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*[73] proposed a hybrid system for automatic recognition of events, temporal expressions and temporal relations in clinical records. Authors combined different ML methods[57] 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.

Personal data anonymization. The protection of personal data has become a challenge for many health institutions, specially with the rising computerization of almost any clinical record. *Ferrández et al*[34] 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[57] 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, when the proposal is compared to the best system, an improvement in more than a 26% for the F2-measure metric. *Benton et al*[7] 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[57]. A CRF[49] 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 versus 0.730).

Relationship extraction. The discovery of semantic relationships, e.g. connections among diseases and symptoms, between medical concepts is essential. *Uzuner et al*[82] discovered different relationships among several clinic entities in a set of hospital discharge reports. Based on the semantic types defined in the UMLS Metathesaurus[9], the following relations were analysed: disease-treatment, disease-test, and disease-symptom. Authors used a ML-based approach that included a SVM[91]. 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*[69] used some electronic medical records to analyse eight associations between medical problems, treatments, and tests. To carry out this task, authors used a ML-based approach (CRF[49] and SVM[91]). They used CRF[49] to extract medical concepts, and different SVM[91] 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*[94] 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 analysed: treatment-problem, test-problem, and problem-problem. To carry out this task, authors used a hybrid approach, based on ML[57], 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[26] (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.

Hypotheses generation and knowledge discovery

The discovery of new hypotheses and hidden knowledge on textual data is essential to provide healthcare professionals with important insights that can be used in their daily practices, and it also supports their research works. The extraction of such valuable knowledge is essential to detect risk factors, symptoms and critical events of a patient, and facilitating, therefore, the arduous task of decision-making that is daily carried out by health professionals.

Discovery of new hypotheses. *Baron et al*[6] performed a meta-analysis to identify adverse effects of aspirin usage, considering TM techniques[40] to select the most relevant articles —3,916 from a total of 119,310 references were taken. To carry out this task authors created an automatic TM procedure to score references of potential relevance to the meta-analysis based on the occurrence of words in the title, abstract and indexing terms. It was discovered that serious gastrointestinal events were very rare, but the use of aspirin was associated with a higher risk of minor gastrointestinal complaints than placebo or active comparators. *Heintzelman et al*[43] studied the importance of combining TM[33], Natural Language Processing (NLP)[19] and UMLS Metathesaurus[9] to properly classify and study different pains of patients with metastatic prostate cancer. Authors applied a multiple regression model to find the existing associations among the occurrence of pain and the rest of the variables defined in the study (e.g. receipt of various drugs). Among other findings, authors discovered that the receipts of opiates and palliative radiation were robustly associated with severe pains, revealing the usefulness of these techniques for the identification of new cancer phenotypes. *Cole et al*[16] considered multivariate logistic regression to analyze clinical notes and to extract associations between allergic conditions and chronic uveitis. In this study, different TM techniques[40] were used including identification of medical entities (e.g.

diseases or drugs), detection of negations, standardization and disambiguation of terms using 22 clinical ontologies for this purpose. Results identified four previously known associations with uveitis, and presented some evidences supporting a new clinical hypothesis such as the association between allergic conditions and chronic uveitis in juvenile idiopathic arthritis patients. *Leeper et al*[50] used TM techniques[40] to detect the adverse events associated with the use of Cilostazol to patients with a peripheral arterial disease. To carry out this objective, different classic tasks of the TM systems were performed, including 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. 1,8 million of subjects from the Standford clinical datawarehose were analysed, demonstrating the non-association between the use of this drug and any major cardiovascular event, e.g. myocardial infarction, stroke or death.

Knowledge discovery. *Tafti et al*[81] 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[57], NLP[19] and distributed processing frameworks such as Apache Spark. The proposal, named Word2Vector algorithm[58], is an algorithm based on Deep Learning[78] that was evaluated obtaining a superior performance (Precision 0.936 and Recall 0.930) than traditional models (e.g. BOW[93]+Decision tree[77] 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*[10] presented a hybrid system, which is based on rule-based NLP and ML[57], 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 discovery 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 UIMA framework[35] for the preprocessing of clinical reports and text analysis tasks. The overall performance of the proposed system slightly exceeded the standard (0.911 versus 0.898 in F-measure). *Collier*[17] provided an overview of the role played by TM techniques[40] to discover novel information from large-scale text collections, in particular in epidemic detections. Author analysed the Biocaster tool[18], a really interesting web service based on NLP[19] and TM techniques[40], which was able to improve the early detection of outbreaks of infectious diseases only 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*[88] 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[78] approach based on a Convolutional Neural Network model[46]. The method was evaluated against four baseline systems (multi-layer perceptron, SVM, random forest and logistic regression)[48], achieving an improvement of 20% in the macro-averaged F-measure.

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 analysed. Summarization techniques[2] can be grouped into different categories (single or multi-document summarization, text or multimedia summarization, general purpose or domain-specific, among others). Nevertheless, according to *Hahn and Mani*[41], the two main groups for these techniques are extractive and abstractive summarization.

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*[28] presented an extractive and multi-document summarization system integrated in PERSIVAL[54](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 non-professionals). 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*[74] 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[57]. The system comprises three different phases: i) document preprocessing; ii) a bagging meta-learner to extract sentences where the C4.5 algorithm[65] was considered as the base learner; and iii) 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).

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*[37] proposed a methodology based on semantic abstraction to find relevant information about some specific diseases based on PubMed search results. Authors considered the SemRep[67], a NLP[19] tool, to extract entities and relations from textual reports. Additionally, it makes use of the UMLS Metathesaurus[9] and the MetaMap Transfer[56] 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 versus 0.170). *Rindflesch et al*[68] 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

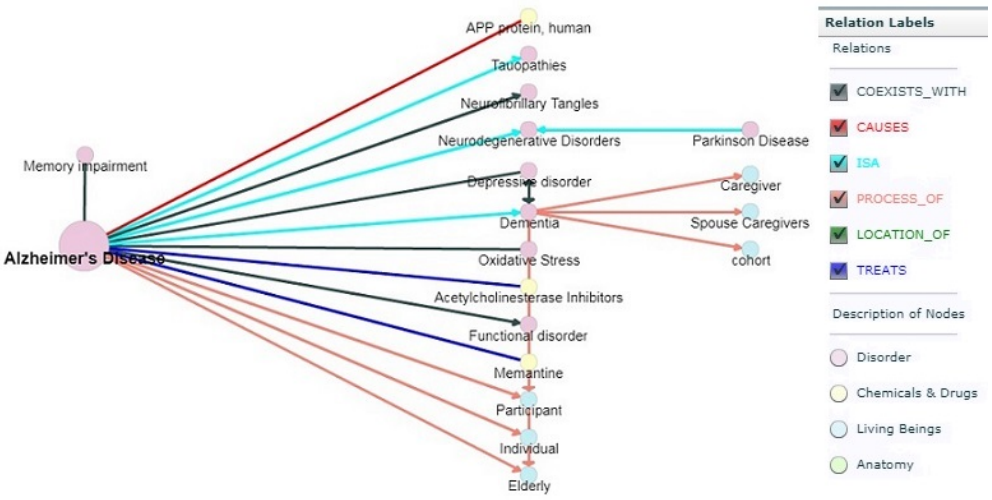


Figure 3: Semantic MEDLINE graph. Example of summarization according to treatment of a disease summary type (query: Alzheimer's disease). The nodes represent different entities found in citations and the arrows the connections between those entities.

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*[92] 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[29] 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 versus 0.470 in F-measure).

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 tools to be used. Some research studies on this matter are analysed, denoting how TM[40] and NLP techniques[19] may automatically produce corpus, ontologies and specialized databases.

Corpus construction. *Roberts et al*[71] described the construction of a semantically annotated corpus of clinical texts, known as CLEF[70] corpus (`http://nlp.shef.ac.uk/clef/`). The annotation methodology is based on NLP[19]. The corpus consisted on structured records and unstructured documents related to 20,234 different patients that 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[22] and UMLS[9] 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*[83] 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 inter-relationships 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.

Ontologies construction. *Fabian et al*[30] 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[40]. 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*[52] employed a statistical TM approach to develop a clinical vocabulary for post-traumatic stress disorder (PSTD) in Veterans Health Administration from outpatient progress notes. Authors used SAS Text Miner[1], 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[25] 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).

Databases construction. *Fang et al*[31] analized 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[3] 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*[87] 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[40]. 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 miRNA-cancer. 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 versus 0.770). According to the authors, the use of TM techniques[40] enabled a minimization of the manual maintenance of the database.

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[60], adverse events classification[42], electronic health records classification[12], symptomatology categorization[84], health miner (opinion and sentiment analysis mining)[5]. Nevertheless, most of research studies are focusing on three main applications: automatic diagnostic classification; patient stratification; and classification of medical literature.

Automatic diagnostic classification. *Metais et al*[55] 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[40] and ML[57] 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 multi-label classification (each clinical report can be identified with multiple diagnostic codes) by including a novel multi-label 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, e.g. stemming by using an adaptation of the Porter[64] 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*[39], 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: i) a first one based on Lucene (`https://lucene.apache.org/core/`)[8] (an open source library that allows analysis of textual collections); ii) another based on BoosTexter[76] (a boosting algorithm for text classification); and iii) 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).

Patient stratification. *Zuccon et al*[95] 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[77], Naive Bayes, SMO and SPegasos, a variation of SVM, to carry out this task. They considered some external tools, e.g. SNOMED CT Thesaurus[25], 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*[44], authors presented a multi-class classification system to automatically identify smoking status (current smoker, past smoker, past or current smoker, non-smoker and unknown status) from unstructured electronic health records. The proposal included different NLP techniques[19] such as stop words removal, tokenization, and stemming. Authors used a hybrid approach using rule-based and ML techniques[57]. 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).

Medical literature classification. *Frunza et al*[38] 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[57] and three types of text representations: BOW[93], UMLS concepts, and a combination of both. The dataset 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*[89] 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[19] and an approach based on SVM[91] 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[93], bag-of-phrases[27] 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).

Text Mining Resources for Medicine

All the existing TM techniques in Medicine are required to be complemented with different resources and tools such as corpus, metathesauruses, ontologies, 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[33] based system, 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`.

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[33] in Medicine. It is a corpus specifically developed to support the construc-

tion and evaluation of IE and TM[33] 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, e.g. disease named entity recognition.

As for the ontologies used in the field of TM and Medicine, *Disease Ontology (DO)* (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 (OAE)* (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 (Unified Medical Language System)* (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 MedicineClinical Terms)* (http://www.snomed.org/snomed$-$ct) is a multilingual clinical healthcare terminology that includes three types of component: concepts, descriptions and relationships.

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 segment into sentences 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 so 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 multi-language processing, including named entity detection, PoS-tagging, parsing, disambiguation, among others.

Named Entity Recognition 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 API whose main objective is the automatic identification and annotation of biomedical concepts (disorders, anatomical concepts, genes, biological processes).

Advanced Text Mining 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 as terminological and semantical basis the UMLS metathesaurus[9]. MetaMap includes different NLP techniques[19], and it is able to perform multiple tasks such as detection of negated terms, words disambiguation or acronyms detection. *UIMA (Unstructured Information Management Architecture)* (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.

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[40] in the clinical field can be summarized as follows: improvement in the quality of care services[24]; reduction in the number of medical errors[15]; supporting the prevention and detection of diseases[45]; identification of the most beneficial treatments[61]; and improving both the time and costs related to the health management[80]. 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[40] as described below.

Advanced systems for clinical decision support. Clinical decision support systems are the result of the synergy of disciplines such as TM, ML[57] 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[36] (`https://www.ibm.com/watson/health/`) is a clear example of how the synergy of various disciplines such as TM, NLP[19] and ML[57] can help in building cognitive systems that support the healthcare 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 healthcare 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*[63], 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[40] 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.

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. Thanks to 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[40]. As an example of this applicability, let us consider the analysis of tweet contents, which provide interesting health behavioral profiles from a population. *Yoon et al*[90] considered the physical activity as the main topic to be analysed within tweet contents, denoting the healthy habits to help in the prevention of diseases, e.g. cardiovascular problems. In a similar way, *Paul and Dredze*[62] 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 et al*[21] provided a disease surveillance resource that was able to identify diseases in online communities. Authors used TM techniques[40] 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[20] algorithm, a graph-based data mining model to identify anomalies in flu from blogs.

Conclusions

This paper has provided an extensive review about TM techniques, performing an analysis and description of more than ninety 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 this information has 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 varied set of TM resources for Medicine have been denoted so they can complement the techniques previously analysed. Finally, different open challenges in clinical medicine have been described and analysed.

Acknowledgements

This work was Supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund, under the projects TIN2014-55252-P and TIN2017-83445-P.

References

[1] Abell M (2014) SAS Text Miner. CreateSpace Independent Publishing Platform, USA

[2] Afantenos S, Karkaletsis V, Stamatopoulos P (2005) Summarization from medical documents: a survey. Artificial intelligence in medicine 33(2):157–177

[3] Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:170702919

[4] Aronson AR, Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17(3):229–236

[5] Asghar MZ, Qasim M, Ahmad B, Ahmad S, Khan A, Khan IA (2013) Health miner: opinion extraction from user generated health reviews. International Journal of Academic Research 5(6)

[6] Baron JA, Senn S, Voelker M, Lanas A, Laurora I, Thielemann W, Brückner A, 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

[7] Benton A, Hill S, Ungar L, Chung A, Leonard C, Freeman C, Holmes JH (2011) A system for de-identifying medical message board text. BMC Bioinformatics 12(3):1–10, DOI 10.1186/1471-2105-12-S3-S2, URL http://dx.doi.org/10.1186/1471-2105-12-S3-S2

[8] Białecki A, Muir R, Ingersoll G, Imagination L (2012) Apache lucene 4. In: SIGIR 2012 workshop on open source information retrieval, p 17

[9] Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 32(suppl_1):D267–D270

[10] Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF (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

[11] Carrero F, Cortizo JC, Gómez JM (2008) Building a Spanish MMTx by using Automatic Translation and Biomedical Ontologies. In: International Conference on Intelligent Data Engineering and Automated Learning, Springer, pp 346–353

[12] Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, Cai T, Hoffnagle AG, Dai Y, Block S, et al (2015) Validation of electronic health record phenotyping of bipolar disorder cases and controls. American Journal of Psychiatry 172(4):363–372

[13] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG (2001) A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics 34(5):301–310

[14] Chen H, Fuller SS, Friedman C, Hersh W (2005) Knowledge management, data mining, and text mining in medical informatics. In: Medical Informatics, Springer, pp 3–33

[15] Cohan A, Fong A, Ratwani RM, 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, ACM, pp 52–59

[16] Cole TS, Frankovich J, Iyer S, LePendu P, Bauer-Mehren A, Shah NH (2013) Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for ehr-based research. Pediatric Rheumatology 11(1):45

[17] Collier N (2012) Uncovering text mining: A survey of current work on web-based epidemic intelligence. Global public health 7(7):731–749

[18] Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo QH, Dien D, Kawtrakul A, Takeuchi K, et al (2008) BioCaster: detecting public health rumors with a web-based text mining system. Bioinformatics 24(24):2940–2941

[19] 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(Aug):2493–2537

[20] Cook DJ, Holder LB (1994) Substructure discovery using minimum description length and background knowledge. Journal of Artificial Intelligence Research 1:231–255

[21] Corley CD, Cook DJ, Mikler AR, Singh KP (2010) Text and structural data mining of influenza mentions in web and social media. International journal of environmental research and public health 7(2):596–615

[22] Cunningham H, Maynard D, Bontcheva K, Tablan V (2002) A framework and graphical development environment for robust NLP tools and applications. In: ACL, pp 168–175

[23] Dandapat S, Way A (2016) Improved named entity recognition using machine translation-based cross-lingual information. Computación y Sistemas 20(3):495–504

[24] 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

[25] Donnelly K (2006) SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics 121:279

[26] 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

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

[28] Elhadad N, Kan MY, Klavans JL, McKeown K (2005) Customization in a unified framework for summarizing medical literature. Artificial intelligence in medicine 33(2):179–198

[29] Erkan G, Radev DR (2004) Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22:457–479

[30] Fabian G, Wächter T, Schroeder M (2012) Extending ontologies by finding siblings using set expansion techniques. Bioinformatics 28(12):i292–i300

[31] Fang YC, Huang HC, Chen HH, Juan HF (2008) TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. BMC complementary and alternative medicine 8(1):58

[32] Feldman K, Hazekamp N, Chawla NV (2016) Mining the Clinical Narrative: All Text are Not Equal. In: Healthcare Informatics (ICHI), 2016 IEEE International Conference on, IEEE, pp 271–280

[33] Feldman R, Sanger J (2007) The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press

[34] Ferrández O, South BR, Shen S, Meystre SM (2012) A hybrid stepwise approach for de-identifying person names in clinical documents. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, pp 65–72

[35] 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

[36] Ferrucci D, Levas A, Bagchi S, Gondek D, Mueller ET (2013) Watson: beyond jeopardy! Artificial Intelligence 199:93–105

[37] Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC (2009) Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. Journal of biomedical informatics 42(5):801–813

[38] 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

[39] Goldstein I, Arzumtsyan A, Uzuner Ö (2007) Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2007, p 279

[40] Gupta V, Lehal GS, et al (2009) A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence 1(1):60–76

[41] Hahn U, Mani I (2000) The challenges of automatic summarization. Computer 33(11):29–36

[42] Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, Jung K, LePendu P, Shah NH (2014) Text mining for adverse drug events: the promise, challenges, and state of the art. Drug safety 37(10):777–790

[43] Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, Childs LC, Bova GS (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

[44] Jonnagaddala J, Dai HJ, Ray P, Liaw ST (2015) A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. ACL-IJCNLP 2015:147–151

[45] 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

[46] Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:14085882

[47] Kipper-Schuler K, Kaggal V, Masanz J, Ogren P, Savova G (2008) System evaluation on a named entity corpus from clinical notes. In: Language resources and evaluation conference, LREC, pp 3001–3007

[48] Kotsiantis S (2007) Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249–268

[49] Lafferty J, McCallum A, Pereira F (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, pp 282–289

[50] Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH (2013) Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PloS one 8(5):e63,499

[51] Lin YK, Chen H, Brown RA (2013) MedTime: A temporal information extraction system for clinical narratives. Journal of biomedical informatics 46:S20–S28

[52] 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

[53] McCallum AK (2002) Mallet: A machine learning for language toolkit

[54] McKeown KR, Chang SF, Cimino J, Feiner S, Friedman C, Gravano L, Hatzivassiloglou V, Johnson S, Jordan DA, Klavans JL, 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, ACM, pp 331–340

[55] Metais E, Nakache D, Timsit JF (2006) Automatic classification of medical reports, the cirea project. In: Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey, pp 354–359

[56] Meystre S, Haug PJ (2005) Evaluation of medical problem extraction from electronic clinical documents using MetaMap Transfer (MMTx). Studies in health technology and informatics 116:823–828

[57] Michalski RS, Carbonell JG, Mitchell TM (2013) Machine learning: An artificial intelligence approach. Springer Science & Business Media

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

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

[60] Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL, et al (2007) Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 13(6 Part 1):281–288

[61] 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

[62] Paul MJ, Dredze M (2012) A model for mining public health topics from twitter. Health 11:16–6

[63] Portela F, Cabral A, Abelha A, Salazar M, Quintas C, Machado J, Santos M (2014) Knowledge acquisition process for intelligent decision support in critical health care. Healthcare Administration: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications 270

[64] Porter MF (1980) An algorithm for suffix stripping. Program 14((3)):130–137

[65] Quinlan JR (2014) C4. 5: programs for machine learning. Elsevier

[66] Ramshaw LA, Marcus MP (1999) Text chunking using transformation-based learning. In: Natural language processing using very large corpora, Springer, pp 157–176

[67] Rindflesch TC, 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

[68] Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D (2011) Semantic MEDLINE: An advanced information management application for biomedicine. Information Services & Use 31(1-2):15–21

[69] 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

[70] Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, Kola JS, Roberts I, Setzer A, Tapuria A, et al (2007) The CLEF corpus: semantic annotation of clinical text. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2007, p 625

[71] 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

[72] Roberts K, Harabagiu SM (2011) A flexible framework for deriving assertions from electronic medical records. Journal of the American Medical Informatics Association 18(5):568–573

[73] Roberts K, Rink B, Harabagiu SM (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

[74] 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

[75] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (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

[76] Schapire RE, Singer Y (2000) BoosTexter: A boosting-based system for text categorization. Machine learning 39(2-3):135–168

[77] Sebastiani F (2002) Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1):1–47

[78] Shickel B, Tighe PJ, 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

[79] Skeppstedt M, Kvist M, Nilsson GH, 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

[80] Sun W, Cai Z, Liu F, Fang S, Wang G (2017) A survey of data mining technology on electronic medical records. In: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), pp 1–6, DOI 10.1109/HealthCom.2017.8210774

[81] Tafti AP, Badger J, LaRose E, Shirzadi E, Mahnke A, Mayer J, Ye Z, Page D, Peissig P (2017) Adverse drug event discovery using biomedical literature: a big data neural network adventure. JMIR medical informatics 5(4)

[82] Uzuner O, Mailoa J, Ryan R, Sibanda T (2010) Semantic relations for problem-oriented medical records. Artificial intelligence in medicine 50(2):63–73

[83] Van Mulligen EM, Fourrier-Reglat A, Gurwitz D, Molokhia M, Nieto A, Trifiro G, Kors JA, Furlong LI (2012) The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. Journal of biomedical informatics 45(5):879–884

[84] Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z, Williams BA, Ebadollahi S, Stewart WF, 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

[85] Wang Y, Patrick J (2009) Cascading classifiers for named entity recognition in clinical notes. In: Proceedings of the workshop on biomedical information extraction, Association for Computational Linguistics, pp 42–49

[86] 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. In: CLEF (Working Notes)

[87] 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

[88] 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

[89] Yetisgen-Yildiz M, Pratt W (2005) The effect of feature representation on MEDLINE document classification. In: AMIA annual symposium proceedings, American Medical Informatics Association, vol 2005, p 849

[90] Yoon S, Elhadad N, Bakken S (2013) A practical approach for content mining of Tweets. American journal of preventive medicine 45(1):122–129

[91] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ (2010) Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 10(1):16

[92] Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC (2011) Degree centrality for semantic abstraction summarization of therapeutic studies. Journal of biomedical informatics 44(5):830–838

[93] Zhang Y, Jin R, Zhou ZH (2010) Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 1(1-4):43–52

[94] 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

[95] Zuccon G, Wagholikar AS, Nguyen AN, 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. AMIA Summits on Translational Science Proceedings 2013:300