Computers in Industry 78 (2016) 43–56

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Turning user generated health-related content into actionable knowledge through text analytics services

Paloma Martínez [a,*], José L. Martínez [b], Isabel Segura-Bedmar [a], Julián Moreno-Schneider [a], Adrián Luna [b], Ricardo Revert [a]

[a] Computer Science Department, Universidad Carlos III de Madrid, Spain
[b] MeaningCloud LLC, USA

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form 2 October 2015
Accepted 13 October 2015
Available online 10 November 2015

ABSTRACT

In the last years, the habit of discussing healthcare issues with family and friends, even with unknown people, in the context of social networks has increased and processing user generated content has become a new challenge. This can help in on-line crowd surveillance for different applications (pharmacovigilance and filtering health contents in blogs among others) as well as extracting knowledge from unstructured text sources. In this article, a system that monitors health social media streams is described. It is based on several text analytics processes supported, among others, by MeaningCloud, a commercial platform which provides meaning extraction from texts in a Software as a Service mode. In this architecture, several domain resources are integrated to detect drugs and drug effects such as CIMA (official information about authorized drugs in Spain maintained by the Spanish Agency of Medicines and Health Products), MedDRA (Medical Dictionary for Regulatory Activities) and the SpanishDrugEffectDB database that contains relations between drugs and effects. Different ways of visualizing data considering time lines and aggregated data have been implemented. In order to show performance, an evaluation has been carried out over Named Entities Recognition (NER) and Relation Extraction (RE) tasks.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Current definitions of Social Media [17] include several sources of user generated data, from Twitter to specialized blogs, through Facebook. Users of these Web 2.0 applications share information about any subject, including issues related to their health condition. The number of people with Internet access seeking for health information through the net ranges from 70% to 75% in the U.S. Besides, 42% of them used social media to get information about health issues. Moreover, mobile technology creates an ecosystem where people are continuously accessing to the Internet and this changes the way people interact with healthcare professionals.

In this context, there is an increasing volume of digital interaction that produces a big stream of data with meaningful information that companies and organizations need to access. In networks and forums such as PatientsLikeMe[1], DailyStrength[2] or Saluspot[3] patients talk to each other about their feelings regarding a health problem, the way their bodies react to a given drug, how they mix different drugs to fight against a specific disease they have and many other issues related to their health condition. They can access health-related content as well as connect and collaborate with other patients looking for health issues.

As an example of the importance of Social Media interactions in the health sector, according to a study developed by Price Waterhouse Coopers, 45% of consumers said information from Social Media would affect their decisions to seek a second opinion[4]. Age distribution of social media users must also be considered; there are surveys indicating that 89% of 18–29 year olds use social media in contrast with the 43% of people aged 65+. The difference of utilization by age groups will diminish over the next years and decades as digital natives increase their involvement and influence professionally and privately within their networks.

In order to analyze this market, the heavily regulated environment around health companies and prevention of direct-to-patient interactions must be taken into account, especially in

* Corresponding author. Tel.: +34916249454.
E-mail address: pmf@inf.uc3m.es (P. Martínez).

[1] http://www.patientslikeme.com/.
[2] http://www.dailystrength.org/.
[3] https://www.saluspot.com/.

[4] PricewaterhouseCoopers, *Social media "likes" healthcare*, http://www.pwc.com/us/en/health-industries/publications/health-care-social-media.jhtml.

http://dx.doi.org/10.1016/j.compind.2015.10.006
0166-3615/© 2015 Elsevier B.V. All rights reserved.

Europe. This prevents pharmaceutical companies to get involved in social networks campaigns and only half of the top 50 pharmaceutical companies in the world interact with patients through social networks. It is also worth mentioning that, outside the U.S., there are a lot of regulatory restrictions forcing pharmaceutical companies to behave in a conservative way. Nevertheless, the interest in listening patients' opinions through social networks as a first step through bidirectional communication with patients is increasing.

Among all health issues, ADRs (Adverse Drug Reactions) are an important health problem due to the fact that they are the 4th cause of death in hospitalized patients [30]. Thereby, the area of pharmacovigilance has captured special attention because of the elevated and increasing frequency of drug safety events [4] along with their high associated costs [28]. Medicine regulatory agencies such as the US Food and Drug Administration (FDA) require clinicians to report every suspected ADR due to the fact that many of them are not spotted in the course of clinical trials. Nevertheless, studies such as [20] claim that ADRs are under-estimated considering that they are communicated by voluntary reporting systems.

Patients can report ADRs using Web-based spontaneous reporting systems (SRS) implemented by medicines agencies such as EMA (European Medicines Agency) and FDA. These SRS have different structures and contents and almost all of them are based on voluntary reporting, except for pharmaceutical companies, which are required to report suspected adverse events once they come to their attention. These companies report adverse drug events to the FDA when there is an identifiable patient, reporter and suspect drug. However, these requirements are not applied to social media.

Patient reports often provide more detailed and explicit information on ADRs than the ones from healthcare specialists [16]. They usually offer a wider or complementary view of the ADR and its possible impact on the patient. Another benefit of patient reporting is that adverse effects caused by OTC (over-the-counter, medicines that are sold without prescription) medications could be analyzed. An important contribution of SRS is getting patients to have a more central role in their treatments. However, despite the fact that these systems are well-established, the rate of spontaneous patient reporting is very low probably because many patients do not know them and may feel either confused when describing their symptoms or even unable to describe them.

On the other side, every medicine is carefully monitored after it is placed on the market, but there are some special drugs, labeled with a *black triangle*[5], that are intensively controlled. This is due to the lack of information available about these medicines compared to others, for example, because they are new in the market or there are few data about its long-term use. In this context, it is therefore essential that the safety of all medicines continues to be monitored while they are in commercial use and that suspected ADRs are reported in order to keep up to date drug packages inserts corresponding to these drugs. Currently, this pharmacovigilance work is carried out by domain experts on a manual basis, by analyzing scientific literature as well as clinical trials documents and spontaneous reports.

Harpaz et al. [14] remarked that new methods that integrate data extracted from SRS narratives and knowledge extracted from experimental preclinical discovery drugs sources are required. Furthermore, patient-generated content concerns also discussions about treatments and opinions about drugs that could lead to valuable knowledge. Patients use Social Media to self-report adverse drug events three times more than reporting to FDA [10] and 90% is the estimated rate of ADRs that patients do not report [23]. Thus, it is reasonable to think that health-related social media can be used as a complementary data source to collect ADRs as well as data about the incorrect use of drugs. In other words, monitoring the abuse and misuse of medicinal products, for instance by people who have problems understanding medical language.

Transforming health-related social media streams into useful knowledge by extracting information from messages is not a trivial issue and requires sophisticated tools to tackle challenges with the overall objective of protecting public health. The two main challenges are (1) to analyze patients sharing experiences and (2) to manage highly informal patient-oriented language, something difficult to deal with as there are barely any resources regarding it.

To cope with these information challenges, Natural Language Processing (NLP) technology is a key aspect and should integrate usable tools to deliver real-time insights to decision makers about surveillance and pharmacovigilance tasks. In this context, this contribution describes the application of text analytics processes to extract information from these real-time Social Media streams relevant for the healthcare sector and the challenges that must be faced. The information to be extracted goes beyond drugs and diseases mentions to show also relationships among medications and ADRs and indications. It also covers trend evolution of those named entities mentions and ADRs detected in patients' conversations.

Text analytics processes to be applied for this purpose cannot be generic, but adapted to the health domain. This requires specific dictionaries and ontologies covering drugs, diseases, body part names and other topics to be integrated in named entity recognition processes and to cope with colloquial expressions and laypeople terms for drugs, diseases and other entities with the aim to discover what is being talked about.

The work presented in this article integrates different semantic resources and processes in a complete framework to face real-time text analytics on social media, in particular for monitoring drug-related medical events (ADRs, indications, symptoms, diseases,...) in online Spanish social networks about healthcare, being able to process a large volume of data in real-time[6] and to address the abovementioned challenges. This system is a result of TrendMiner European project.

Remainder of the article is organized as follows. Section 2 is devoted to review the state of the art in drugs and ADRs entity recognition in medical literature and social media. Section 3 describes the functionality of the proposed system, its architecture and integrated semantic resources. Section 4 describes the evaluation carried out and finally, conclusions and future work are given in Section 5.

2. Related work

Due to the fact that users are active consumers and producers of health-related contents in Internet, in the last years extracting knowledge from unstructured contents (mainly texts) in health domain has received a great attention. The main reason is these sources could reveal important public health issues. Many efforts have been devoted to the application of NLP techniques to gather information about health issues, such as diseases, symptoms, drugs, adverse events and others from texts. A comprehensive overview of the application of text mining techniques to biomedical knowledge extraction from scientific literature, clinical narratives and on-line health web sites is given in [35].

Focusing on analysis of social media to mine data about personal health, there are many works that use Twitter both to detect pre-established health conditions and unknown trends. Parker et al. [22] described a method to identify emerging public

[5] http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/document_listing/document_listing_000365.jsp.

[6] The system can be accessed at http://trendminer.daedalus.es.

health conditions that is based on using a set of frequent term sets extracted from 1.6 million tweets and analyzed with time series to show how prevalent a keyword is over time. These sets are then connected to Wikipedia articles over a series of time windows to show relevance.

In the topic of drugs and ADRs, there are works that use different sources, mainly in the context of drug product labels [13–19], biomedical literature [31], medical case reports [12] and health records [27]. In recent years, there has been an increasing interest in the extraction of ADRs from social media, although it has been investigated to a limited extent

Different recent research works about mining the pharmacovigilance literature are included in [40]. We refer the reader to the article [32] for an excellent survey describing research works that use social media for pharmacovigilance. A relevant work is, for example, Leaman et al. [18] where a system was developed to automatically detect adverse effects mentions in user posts focusing on four drugs that are known to cause ADRs. A corpus annotated with indications and ADRs consisting on 3600 comments extracted from the DailyStrength health-related social network was used for developing and testing purposes. The system obtained a precision of 78.3% and a recall of 69.9%. Nikfarjam and Gonzalez [21] extended the Leaman et al.'s work by adding association rule mining techniques to extract prevalent patterns concerning patient opinions about drug treatments. In this case, a precision of 70.01% and a recall of 66.32% were achieved. The main disadvantage was that patterns were too domain-dependent but this work allows detecting terms that are not included in dictionaries.

Another proposal is Bian et al. [3], which describes the use of SVM classifiers to recognize tweets about ADRs. The classifier is trained on a corpus of tweets labeled with UMLS (Unified Medical Language Systems) concepts using MetaMap tool [1]. The accuracy is low because not only people use informal language to talk about their medical condition but also Twitter has other challenges such as slangs, poor language structures, URLs and emoticons among others. This makes MetaMap inadequate to process this type of texts.

There is also an intensive work creating corpora labeled with entities such as drugs, adverse events, diseases, etc., see [34] for a complete review of corpora. The availability of such resources in NLP community is very important to train and test text mining algorithms. For instance, Benton et al. [2] made a corpus of posts from different online blogs about breast cancer that was used to extract potential ADRs from the four most common used drugs to treat this disease. A lexicon was built from several websites and databases such as Consumer Health Vocabulary (CHV)[7] and integrated in a co-occurrence based approach to detect significant pairs of drug-ADR. To evaluate the system, ADRs from drug labels were collected and precision and recall were calculated by comparing the ADRs from drug labels and the ADRs obtained by the system. The system obtained an average precision of 77% and an average recall of 35.1% for all four drugs. An available corpus created by extracting tweets related to 74 drugs and their variant names is described in [11]. A total of 10,822 tweets are annotated with the presence of ADRs, the span of the ADR mention and its UMLS identifier. Other corpora have been focused on other languages, such as the corpus described in [26], a Spanish corpus of user comments extracted from a health forum that is annotated with drugs and ADRs.

State of the art in drug NER shows performance on scientific texts is around 85% in terms of F-score. The best system in the BioCreative IV (2014) CHEMDNER (Chemical compound and drug name recognition) task achieved an 87.39% of F-score working on Pubmed abstracts, decreasing performance to 60% when informal texts were considering [2]. Regarding ADRs detection in English social media, systems are between 60% and 70%; the use of CHV helps understanding lay terminology but new methods to collect new terms used by patients are required. In the case of Spanish language, there is no such resource and this makes ADRs even more difficult recognition.

Regarding the state of the art in relation extraction tasks in the biomedical domain, in recent years much effort has been devoted to extraction of protein–protein interactions from biomedical literature; [36] reports 58% of F-score with a kernel method using the AIMed corpus. Another important research area is extraction of drug–drug interactions from texts (see [37] for a detailed analysis of DDIExtraction 2013 participating systems). The best system achieved a 53% of F-score over Medline abstracts and 82% of F-Score over DrugBank texts. These works use mainly supervised machine learning methods. Due to the fact that annotating corpora is a highly cost task, unsupervised methods, which do not require annotated data, are being developed. Recently, an alternative paradigm based on distant supervision that uses unlabeled corpora extracted from a patient's blog has been proposed [25]. The distant supervision hypothesis establishes that if two entities appear in a sentence, then both entities might participate in a relation. The learning process is supervised by a database, rather than by annotated texts. In this case, the database used was the SpanishDrugEffectDB database, which relates drugs to their ADRs and indications as is described in Section 3.2. This method achieved an F-score of 53%.

In a different vein, there are other works analyzing patient opinion about drugs such as [6], which demonstrates the ability to track trends in people's positive or negative opinion regarding particular drugs over time. Specifically, an experiment with *Tysabri* drug, indicated for multiple sclerosis, is reported. This work shows that changes in opinion are related to FDA announcements and publicity.

Concerning research projects that cover initiatives for on-line crowd surveillance, EU-ADR Project [8] focused on combining spontaneous reports with electronic healthcare records (EHR) to investigate adverse drug events in Europe. WEB-RADR project [29] was funded by the Innovative Medicines Initiative (IMI) to address the potential of the reporting of ADRs through mobile applications and the recognition of drug safety signals from user comments in social media. TrendMiner project[8] explores Spanish social media (Twitter and forums) to monitor medical events related to drugs [25]. The website Healthmap[9] delivers real-time intelligence on a broad range of emerging infectious diseases for a diverse audience including libraries, local health departments, governments and international travelers by combining online informal sources (online news aggregators, validated official reports, etc.).

Finally, there are companies and institutions that develop APIs (Application Program Interface) that can be integrated in systems. The OpenFDA[10] project provides open APIs, data downloads, and a developer community for high-value public datasets (medical device reports, enforcement reports and drug adverse event reports since 2004 annotated with entities). PatientOpinion[11] is a UK organization that gathers patients' opinions about heath care and treatments they have recently received and offers a read-only API. Novartis company is developing a semantic API to access the data for specific drugs and information about the indications and usage of the drug, its dosage and administration—both for varying patient populations, the known ADRs when taking the drug, known interactions with other drugs and information about clinical studies with the drug.

[7] http://www.consumerhealthvocab.org/.

[8] http://www.trendminer-project.eu/.

[9] Healthmap.org.

[10] https://open.fda.gov/drug/event/.

[11] https://www.patientopinion.org.uk/.

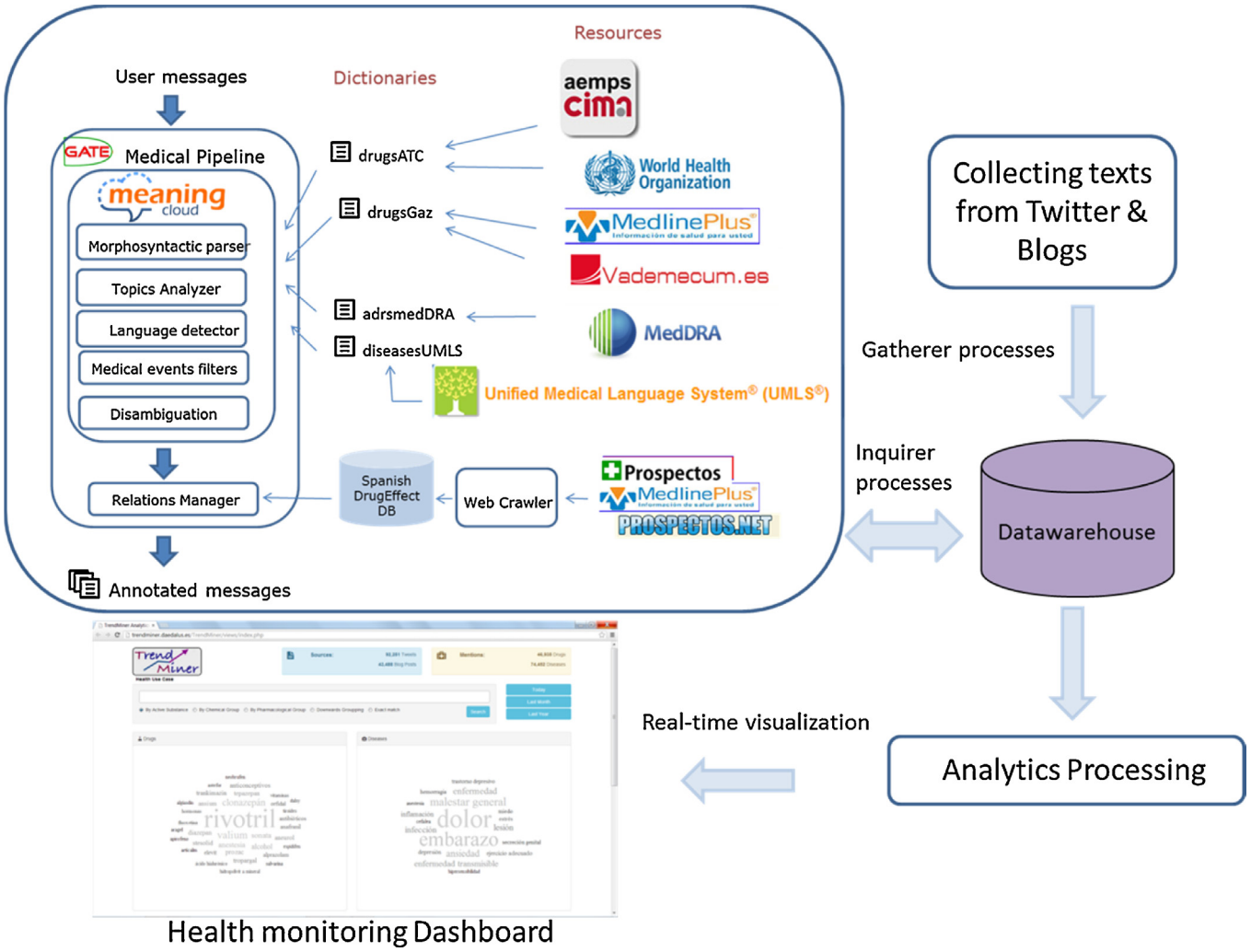

Fig. 1. Health monitoring system architecture.

In summary, there is much room for improvement regarding the analysis of drug's safety and health trends in social media by tracking people messages using NLP technology. Tools that help recognizing patient oriented vocabulary are required in order to understand what people are saying. Moreover, other languages different from English are demanding this kind of tools. This is the focus of the system introduced in this article.

3. System to monitoring health social media

The contribution of this article is a system that monitors patient posts extracted from two Spanish health-related social media. Fig. 1 shows the architecture of this system which is composed of five main components. The central component is the data warehouse, which acts as a core information repository. A set of gatherer processes feeds the system with texts extracted from different sources, while another set of concurrent inquirer processes analyzes the collected texts using a pipeline-based language processor. Finally, the visualization module provides an interface to analyze the data and thus, help discover data insights.

3.1. Annotation flowchart

A set of concurrent inquirer processes uses the GATE[12] Annotation Pipeline that is in charge of specific tasks of annotation and post-filtering (see Fig. 1). The system manages the semantic annotation of the text and the control of concurrency required to deal with such volume of data. There is the possibility to run several annotation processes; the inquirer provides the exclusion mechanisms among the different annotation processes assuring avoiding data corruption and hence, assuring veracity. Each of those processes seeks for the latest created and unlocked text, reserves it by locking it, and then executes an instance of the GATE Annotation Pipeline, which semantically annotates the user comments and tweets. Finally, it stores the response given by the pipeline back to the data warehouse.

MeaningCloud[13] commercial tool is the core technology of this annotation pipeline. It offers several Semantic APIs in SaaS (Software as a Service) mode to extract elements of meaning (topics, facts, opinions, relationships...) from all kinds of unstructured multimedia content. In particular, the functionality integrated in this project has to do with:

- *Topics extraction*[14]: the term *topic* makes reference to named entities and concepts. The algorithm implemented in MeaningCloud to identify named entities is based on hybrid approaches, i.e., the combination of machine learning processes with linguistic ones supported by lexical resources. The ability to customize these lexical resources in a comfortable way is one of the main features of MeaningCloud. Thanks to this, it has been

[12] General Architecture for Text Engineering https://gate.ac.uk/.

[13] http://www.meaningcloud.com/.

[14] https://www.meaningcloud.com/developer/topics-extraction/doc.

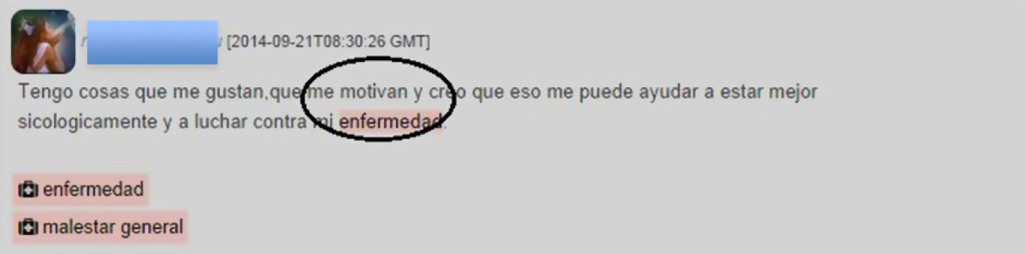

Fig. 2. Example of tweet annotated after the post-filtering stage.

possible to include of clinical and medical vocabulary allowing the detection of drugs, diseases and others. MeaningCloud provides the analysis capabilities needed to deal with different desinences appearing when this vocabulary is used in regular language. An easy example: in Spanish, 'cáncer' can be used in plural, 'cánceres', but both make reference to 'neoplasia maligna'.

- *Syntactical parsing*[15]: if a deep analysis of a sentence is needed, algorithms for part-of-speech tagging and morphosyntactic analysis of texts must be available. MeaningCloud provides this technology for different languages, including English and Spanish. These parsing algorithms combine statistical based approaches and thesauri based ones to produce a deep parsing of sentences.
- *Language identification*[16]: linguistic based techniques are language specific so it is relevant to know in advance in which language a text is written. This is the purpose of this MeaningCloud API, based on a combination of n-gram analysis of text chains and machine learning based classification algorithms.

Improving these Natural Language Processing capabilities was not a goal for this research work. On the contrary, nowadays, NLP technology is a commodity, so commercial products providing this kind of analysis out-of-the-box can be found. The selection criteria followed in this project has been customization capabilities, i.e., easiness in defining the application domain where the NLP algorithms will be applied.

The GATE Annotation Pipeline is composed of six stages:

1 Language identification: stage that discards every text that is not written in Spanish. The gatherer already asks to the Twitter API for texts only in Spanish, but since this may sometimes fail, another filtering step is performed while analyzing the document. The identification is made by the MeaningCloud Language Identification API, which uses statistical techniques based on n-grams.

2 Morpho-syntactic parsing: it is performed by the MeaningCloud Lemmatization, Part-of-speech (PoS) and Parsing API, described above. This API provides morphological information (such as the number and gender of a word) and also syntactic data, including the syntactic category of a word (if it is a name, a verb, an article, ...) as well as the syntactic structure of the sentence (identifying phrases acting as complements, subjects, objects, and so on). The quality of this output depends on the source text type: Twitter messages are written in a way that makes it very difficult to extract a syntactic structure for the text but blog posts are usually well-written from a grammatical point of view. Of course, the output of this analysis is relevant for the semantic disambiguation process, to be performed in a later stage, because it can be useful to distinguish the sense in which a word has been written. For example, 'Motivan' in Spanish is the present form for the third person plural for the verb 'motivar', but it is also a name of a commercial drug. To distinguish which use is being made in a text, morphological information is relevant: if it is tagged as a verb, it cannot make reference to the commercial drug.

3 Topics analyzer: Several health-related dictionaries were created to adapt MeaningCloud Topics Extraction API to the health domain. These dictionaries include drugs, diseases, ADRs, etc. (see Section 3.2). Besides, this component integrates the MeaningCloud topics Extraction API into the analysis process implemented through the GATE pipeline. MeaningCloud provides a plug-in for the GATE platform that makes this integration straightforward. Health related vocabulary is sometimes orthographically complex so it is very easy to make mistakes when writing a drug or disease name. For this reason, this Topics Analyzer uses word matching processes based on the Levenshtein similarity measure that calculates distances between words. For two given words, this distance depends on the number of letters changed, inserted or removed from one of them to obtain the other one.

4 Medical events filter: It filters all the entities that have been annotated by the Topics Analyzer and which are not from the medical domain. Only *drug*, *effect* and *disease* entities are kept in the system.

5 Disambiguation: A set of rules that uses linguistic features like the morpho-syntactic information provided by the parser, together with co-occurrence information of drugs and diseases, is used to filter out terms that are not likely to be mentions of medical events. If we look at the example shown in Fig. 2, *motivan* is a drug name in the drugs dictionary (it is an antidepressant whose active substance is *paroxetina*). However, in this case the syntactical parser detects that it plays as a verb (*motivar*[17]) and consequently it should not be annotated as a drug entity. The rule that is applied in this example is "PoS tag(token_i) ≠ NOUN AND Topic(token_i) = DRUG → delete DRUG topic annotation of token_i". In case that the token is a noun then disambiguation process looks for any other domain-related term in the sentence as is shown in the rule: PoS tag(token_i) = NOUN AND Topic(token_i) = DRUG|ADR|DISEASE AND Topic(token_j) = -DRUG|ADR|CONTEXT WORD → Keep DRUG|ADR|DISEASE topic annotation of token_i (where $j \neq i$ and $1 \leq i, j \leq N$, N is the number of terms in the message being processed and CONTEXT WORD[18] is a medical domain related word different from drugs, ADRs or diseases). These rules have been obtained by manual inspection of a sample of posts.

6 Relations manager: this component annotates three types of relations between drugs and diseases or effects, classifying them into (1) adverse effects, (2) indications or (3) pairs that hold a possible relationship. The two first classifications are relations that were extracted from the SpanishDrugEffectDB database, which has been built from several websites containing drug package inserts as it is explained in Section 3.2. In contrast, the

[15] https://www.meaningcloud.com/developer/lemmatization-pos-parsing/doc.
[16] https://www.meaningcloud.com/developer/language-identification/doc.
[17] to motivate.
[18] Some examples are: *to prescribe*, *to take*, *medication*, *effect*, *produce*, *pill*, *tablet*, ...

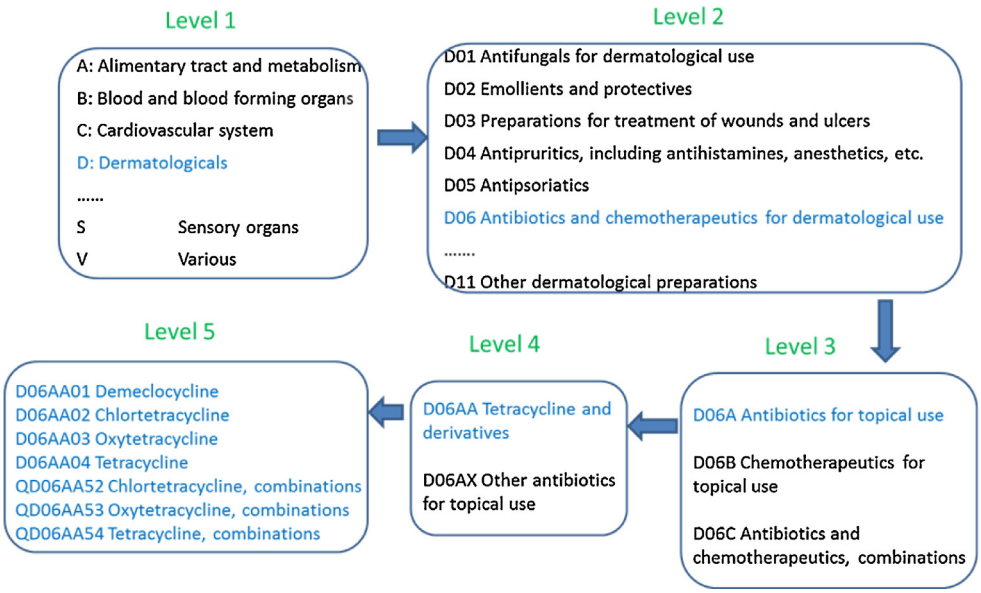

Fig. 3. Example of ATC system structure.

latter group has been created to point out possible un-cataloged or unknown relations that may be discovered due to situations of high recurrence pointed out by the system. For instance, Fig. 6 shows an adverse effect between drug *Taxotere* (Taxotere) and effect *eritema cutáneo* (cutaneous erythema).

3.2. Resources: Drugs, diseases and effects

There are several semantic resources integrated in the system each of them intended to detect a different type of named entity or relations as it is explained below.

3.2.1. CIMA

CIMA is a resource provided and maintained by The Spanish Agency for Medications and Healthcare Products (AEMPS). It is an application which includes all authorized drugs in Spain. The application encloses the following information related to the authorized drugs in Spain: drug's name, active substance(s), marketing authorization holder's name, national code, drug's data sheet, drug's package insert, authorization date, ATC (Anatomical Therapeutic Chemical) code and others. The drug's data sheet is a document which includes the drug's description, indications, dosage, precautions and counter-indications, adverse reactions, pharmaceutical information and properties. The package insert is the document included inside the box of the medicine, and whose goal is to inform the patient.

From CIMA files, 16,418 drugs, 2228 active substances and 3659 brand drugs were obtained. Additionally, 4817 drug related terms were obtained from Vademecum[19] (a guide of pharmaceutical products that includes over 18,200 drugs) and from MedlinePlus[20], the National Institutes of Health's (NIH) website intended for patients. These terms compose the gazetteer *DrugsGaz*.

In order to relate brand names and active substances we use the ATC system that consists on a set of alphanumeric codes developed by the WHO for the classification of drugs and other medical products organized in 5 levels (see Fig. 3). Level 1 represents the part of the body where the drug performs its activity, level 2 represents the therapeutic group, level 3 is about the pharmacological group of the drug, level 4 is the chemical group and level 5 concerns the active substance group. Therefore, it is the key to obtain the relations among drugs and brand names. Wikipedia has a complete and well-structured article dealing with the ATC codes[21] in Spanish that has been crawled to obtain all the existing Spanish ATC codes (4361 ATC codes).

All this knowledge is related to in a dictionary called *drugsATC*. Each entry corresponds to a drug (brand name) followed by those active substances that compose it as aliases. Table 1 shows an example of *dalsy's* dictionary entry, whose id is 896. This drug's composition includes one active substance: *ibuprofeno*. The ATC code associated with the drug *dalsy* is M01AE01. As it is explained below, in the SpanishDrugEffectDB database the ATC codes are also related to the active substances, but this information is not included in the *drugATC* dictionary. In this example, the ATC code of *ibuprofeno* is M01AE01.

Thanks to the information that the ATC provides (therapeutic and chemical characteristics of the drug), the system is able to relate the drugs to each other and is able to categorize them by active substance, chemical group or pharmacological group. This is possible due to the classification's hierarchy. ATC codes are divided in five levels. For example, the ATC code M01AE01 is divided into: Anatomical main group (M), therapeutic main group (01), therapeutic/pharmacological subgroup (A), chemical/therapeutic/pharmacological subgroup (E) and the chemical substance (01), forming the final M01AE01.

3.2.2. MedDRA

MedDRA is the adverse event classification dictionary approved by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), and therefore a very reliable resource for the adverse events.

MedDRA supports ten languages and is composed of a five levels hierarchy which goes from more general to very specific. The two lower levels from MedDRA PT (Preferred Terms) and LLT (Lowest Level Terms) were extracted to implement the *adrsMedDRA* dictionary for ADRs detection. Each LLT is a single medical concept for a symptom, sign, disease diagnosis, therapeutic indication, investigation, surgical or medical procedure, and medical social or family history characteristic. For example, *dolor*

[19] http://www.vademecum.es/.
[20] http://www.nlm.nih.gov/medlineplus/spanish/.
[21] http://es.wikipedia.org/wiki/C%C3%B3digo_ATC.

Table 1
Example of *dalsy* entry in the DrugATC dictionary.

Entry	Drug	Aliases	Morphological tags	Semantic tags	ATC code
896	Dalsy	Ibuprofeno	NPUU-N-dalsy	Sementity/class = instance@type = Top > Drug	SemId_list/ATC = M01AE01

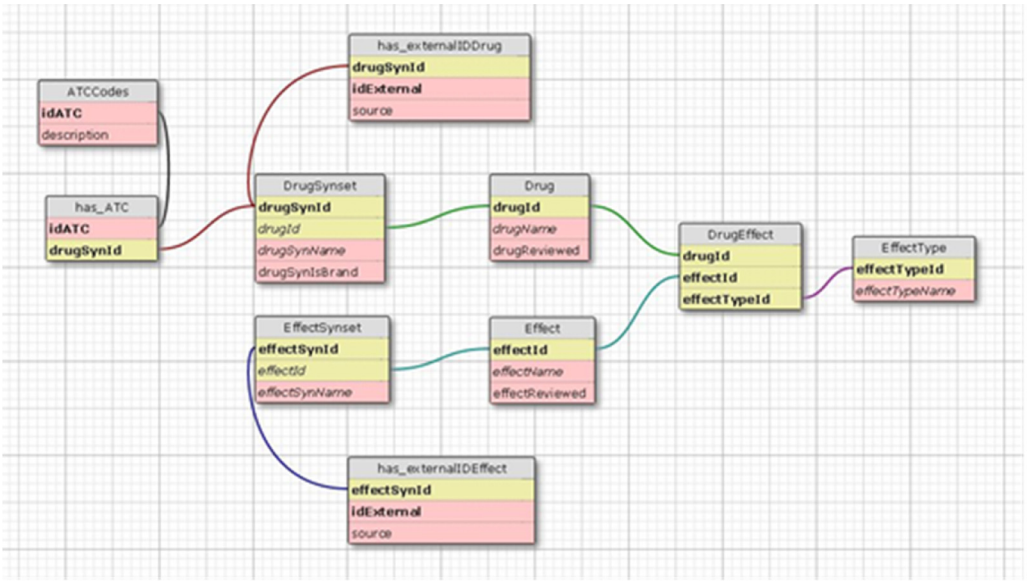

Fig. 4. Class diagram of Spanish drug effect database.

de cabeza (Headache) is a LLT whose PT is the more general adverse event *cefalea* (Cephalalgia). Thus, a relation between general and specific adverse events can be depicted from layers 4 and 5 in MedDRA, and therefore this relation can be used for entries of the dictionary and its aliases. We decided not to include terms corresponding to the *Procedimientos médicos y quirúrgicos* (Surgical and medical procedures) and *Exploraciones complementarias* (Investigations) categories since they do not represent drug effects.

Finally, the information we obtained from this resource is: 13,245 PT adverse effects and 35,259 LLT adverse effects.

3.2.3. UMLS-SNOMED CT

UMLS, developed by the National Library of Medicine (NLM), is a comprehensive list of medical terms mainly focused on developing computer systems suitable for understanding the specific vocabulary which is normally used in biomedicine and health care literature. One of the resources integrated in UMLS is SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms), a terminology accessible in Spanish whose content consists of concepts, descriptions and relationships to represent information and clinic knowledge.

UMLS is structured in several semantic categories (substances, organisms, health care activity, etc.). Three of these categories ('Diseases or syndrome', 'Mental or Behavioral Dysfunction' and 'Neoplastic process') have been chosen in order to create the *diseasesUMLS* dictionary for diseases and symptoms. Some of the extracted terms were PT and others Synonyms. The PTs were set as canonical expressions in the dictionary, and the Synonyms for these were considered as being the aliases.

An extra information field in the entries is the UMLS CUI (Concept Unique Identifier), a code which relates a specific medical term to a set of resources included in UMLS. As a matter of fact, some terms are included in both the diseases dictionary and the adverse effect one. This is the case of *depresión* (depression). In the *adrsMedDRA* dictionary, it is related to its MedDRA code '10012378'. Thus, if checking this code in the UMLS resource, the CUI assigned to it is 'C0011570' (the code included in the *diseasesUMLS* dictionary).

The information we obtained from UMLS Database is 42,548 main diseases and 23,677 diseases synonyms

3.2.4. The SpanishDrugEffectDB database

The last resource used in the annotation pipeline of Fig. 1 is a semantic resource that stores relations between drugs and effects (see Fig. 4). Although there are several English databases such as SIDER[22] or MedEffect[23] with information about drugs and their side effects, none of them are available in Spanish. Moreover, these resources do not include drug indications. There are other initiatives to build knowledge bases in English with ADRs from drug package inserts that can be used to assess ADRs such as [5]. SpanishDrugEffectDB [24] has been built automatically with information about drugs, their indications and their adverse reactions in Spanish. SpanishDrugEffectDB was populated with all drugs and effects from *adrsMedDRA* and *drugsATC* dictionaries.

To obtain the relationships between drugs and their effects, several web crawlers with jsoup[24] parser were developed in order to gather sections describing drug indications and adverse drug reactions from drug package inserts contained in the following websites: MedLinePlus[25], Prospectos.Net[26] and Prospectos.org[27]. Once these sections were downloaded, their texts were processed using the same annotation pipeline of Fig. 1 to recognize drugs and their effects. As each section (describing drug indications or adverse drug effects) is linked to one drug, the effects contained in the section were considered as possible pairs in the relationships with this drug (as

[22] http://sideeffects.embl.de/.
[23] http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index-eng.php.
[24] http://jsoup.org/.
[25] http://www.nlm.nih.gov/medlineplus/spanish/.
[26] http://www.prospectos.net/.
[27] http://prospectos.org/.

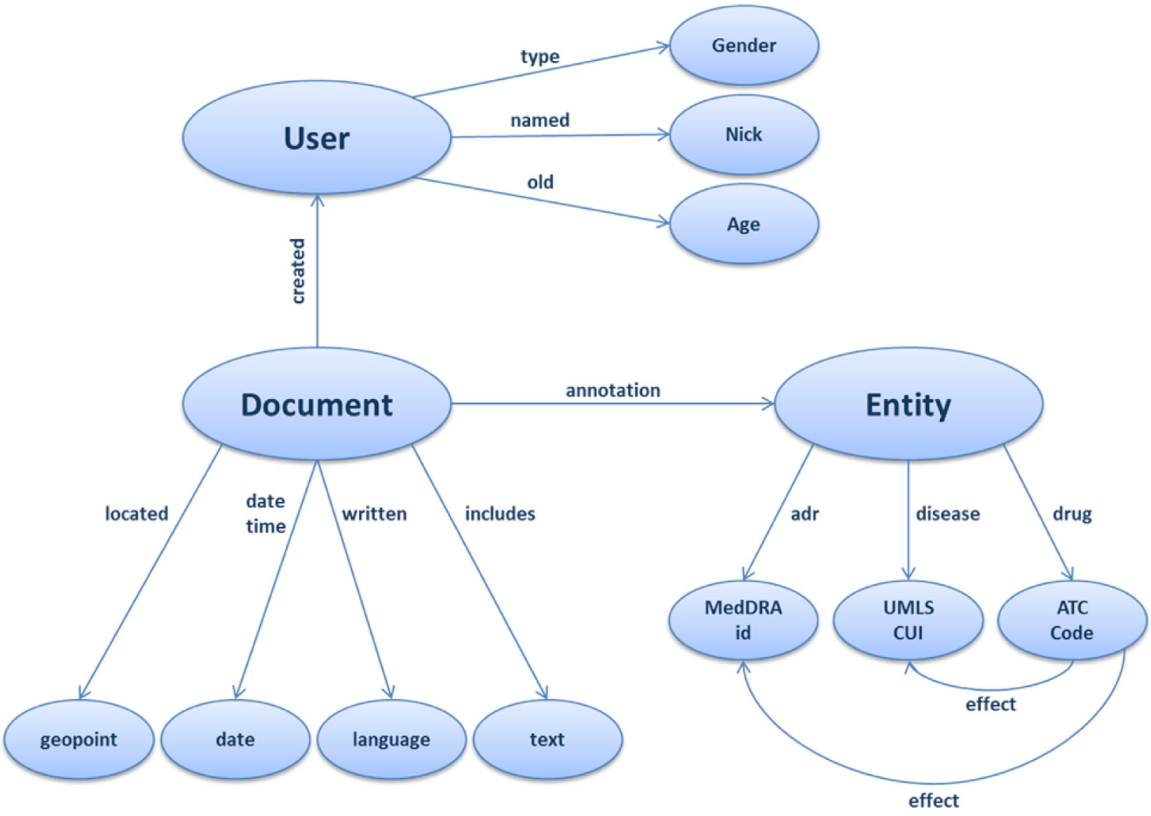

Fig. 5. Datawarehouse index definition.

indication or as adverse drug reaction). More details about this database and a preliminary evaluation can be found in [24].

3.3. Gatherer processes

Two sources of user generated data related to health issues are used: Twitter and Saluspot. From Twitter, tweets that contain specific keywords and which are written in Spanish have been selected. The prototype currently collects tweets containing around 500 antidepressants and related drugs (some examples of keywords are: Clonazepam, Alprazolam, Lorazepam, Rivotril, Clonagin, Diocam, Karidium, ...). The second data source is Saluspot, a Spanish website that allows its users to address free of charge and anonymously their doubts and information needs about health, lifestyle and drugs to thousands of registered doctors. Once a question is posted any of the registered, accredited doctors can answer and even multiple answers are possible. Each question contains information about the user's gender and age, the date of posting and one or more answers together with the identity of the doctor who answered and a reliability measure based on the number of doctors who accepted to tackle this particular question.

Two different subtypes of gatherer processes regarding the type of data source they are querying have been deployed. On the one hand, a process to query the Twitter APIs has been developed in order to collect tweets compliant to the filter. These Twitter-gatherer processes work in real time, feeding the system with new tweets. In order to do so, the Twitter Streaming API[28] is used.

On the other hand, a crawling process that collects posts from Saluspot has been developed. However, due to the intrinsic complexity of the crawling process and the lack of an API that would have been able to alert our system, notifying that new posts (or new answers to the already collected posts) are available, this Saluspot collection does not work in real-time, but collects posts as a snapshot of the currently available information in this website. Also, in this first version, only the most reliable answer was collected.

3.4. Datawarehouse management

A data warehouse based on Elasticsearch[29] is responsible for efficiently storing the high volume of real-time data from social networks that the system manages, as well as for providing advance search functionalities that allow the visualization module to generate complex analytics. Elasticsearch is a flexible, powerful, open source, distributed and real-time search and analytics engine. Some of the key factors that made us take the decision of choosing this datawarehouse are: its distributed capabilities and the fact that it can be easily and horizontally scaled when the system growth starts affecting performance. There are valid Relational Database Management System (RDMS) alternatives but Elastic-search was preferred because it runs on top of Apache Lucene, so it offers quite complex search capabilities, high-performance and is trustworthy, due to its well-known reliability. Using a RDMS for this application was not under consideration due to data type and real-time processing requirements: millions of tweets must be gathered and stored in real time, supporting real-time textual queries; Elasticsearch provides good rates at insertion, scrolling data sets and aggregations tasks.

Fig. 5 shows a graphical representation of the index created to manage the datawarehouse. Documents are in JSON format, they can be tweets or user comments and they are annotated with user features (nickname, age and gender) and other metadata such as geolocation, date, language and the text it contains. Entities can be drugs (represented by ATC codes), diseases (represented by UMLS

[28] https://dev.twitter.com/overview/documentation.

[29] http://www.elasticsearch.org/.

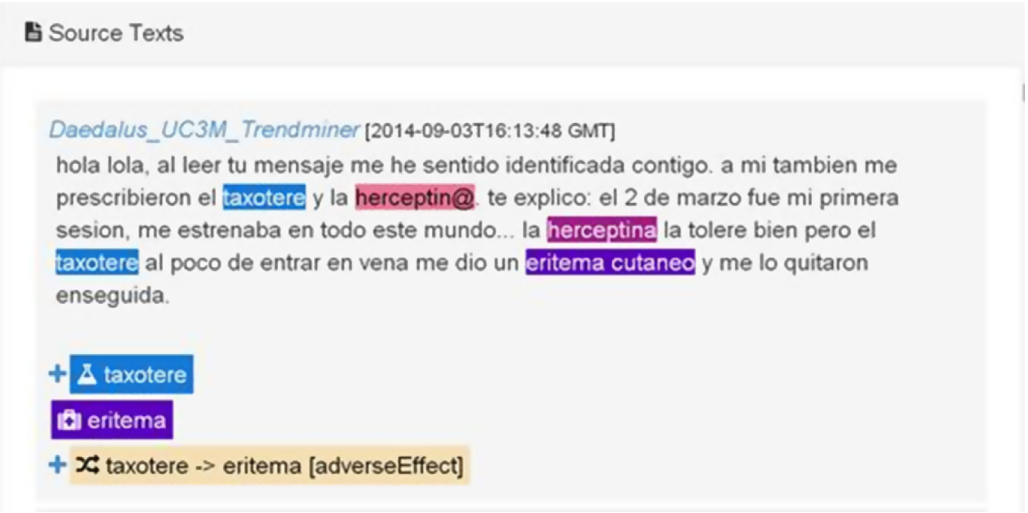

Fig. 6. Example of a comment tagged with drugs, effects and relations.

CUI), and ADRs (represented by their corresponding MedDRA code). Moreover, ADRs and diseases can be related to drugs by effect relations that could be indications, adverse effects and other possible relationship. See Section 3.2 for more details about these concepts and relations.

The index has currently[30] a size of more than 2.5 million entries (documents) that comprises about 1.3 GB of data. The system has been collecting data starting from 17th July 2014 and is up right now. Some statistics from the documents stored in the datawarehouse, at the moment of writing this article are presented next.

Regarding Saluspot source, 41,991 documents were extracted. Every document has already been processed by the GATE pipeline. Within this dataset the system achieves to annotate 1864 mentions of unique drugs as well as 1581 of unique diseases and 2089 of unique adverse effect mentions. Also, regarding the relations extracted, 18,397 unique relations where annotated in this dataset distributed as 1987 adverse effects, 459 indications and 15,951 uncategorized relations (possible relation).

In regards to the Twitter dataset, the system has around 2,760,628 tweets containing a total of 2428 unique drugs, 1681 unique diseases and 2200 unique adverse effects[31]. Also, 25,313 unique relations were extracted, distributed as 97 adverse effects, 570 indications and 22,221 possible relations.

These statistics shows that specific sources for the medical domain (like Saluspot) contain a much higher variability not only regarding to the terminology, but also to the concepts mentioned, as well as a higher number of mentions of medical related concepts per document, when compared to general purpose streams like Twitter. However, some of the drugs, diseases and adverse effects detected in the Twitter stream where not present in the Saluspot dataset, thus confirming the importance of this kind of massive data streams as a source of knowledge.

3.5. Dashboard to visualize monitoring data

Finally, the Analytics Processing component performs the calculations in order to display meaningful relations, patterns of co-occurrence, and other data insights to the final user. The interface provides several search modes based on the drugs ATC code. Three search modes are built around level 3 (Pharmacological Main Group), level 4 (Chemical Main Group) and level 5 (Active Substance) of the ATC structure and they are based on grouping mentions that share the same group at each corresponding level. For example, Fig. 8 shows the result of searching *Trankimazin* drug by *Active Substance* search mode where drugs that share the same active substance (*Alprazolam*) are displayed as well as co-occurrences of these drugs with effects.

Downwards Grouping search mode focuses on the element searched to decide up to what level it should group. It basically gets the element that defines the search, which actually is a node (or several nodes) in the ATC structure tree, and groups together every element below this node or nodes. Finally, *Exact Match* looks for specific mentions of the terms regardless of the ATC code of the mentions. It also includes a fuzzy matching based on the Levenshtein distance to overcome misspelling errors.

The prototype allows viewing the annotated source texts that match a specific search, focusing on their drug and disease mentions and showing the discovered relations, like shown in the example[32] of Fig. 6.

In order to enhance usability, the search box is designed to display every possible term in our vocabulary. The resources used to build the different dictionaries are also compiled and indexed into another Elasticsearch index, using an n-gram analyzer at index time (using from 2 to 20 grams for each word indexed). In contrast, at search time, standard tokenizer, lower case token and stopwords filters are applied. By doing this, the system quickly responds to the user with the hundred most likely terms that match the input provided. Another advantage of this approach is that it allows looking for both the canonical form and any of the synonyms or alias that the term has in the database. See Fig. 7 for more details.

Individual bar graphs aggregating the number of mentions of discovered relations for the texts that match the query are presented, as well as drugs co-occurring with the search term and diseases mentioned along with the search term. Fig. 8 shows mentions of *lorazepam* active substance and brand names that contain it (such as *Orfidal* and *Donix*). Furthermore, we can see their corresponding indications in green (*anxiety* and *insomnia*), ADRs in orange (*depression*, *tremors*) and unknown relationships in blue (*fear*, *stress*). Fig. 9 shows mentions of drug–drug pairs and drug–symptoms pairs.

[30] By july 30, 2015.
[31] By July 30, 2015.

[32] English translation of comment: *hi lola, reading your post I identified myself with you. I was prescribed taxotere and herceptin@ too. I will explain myself: March 2nd was my first session, I debut in this whole world... I tolerated herceptin good enough but soon after taxotere entered my veins it caused me a cutaneous erythema and immediately they told me to stop taking it.*

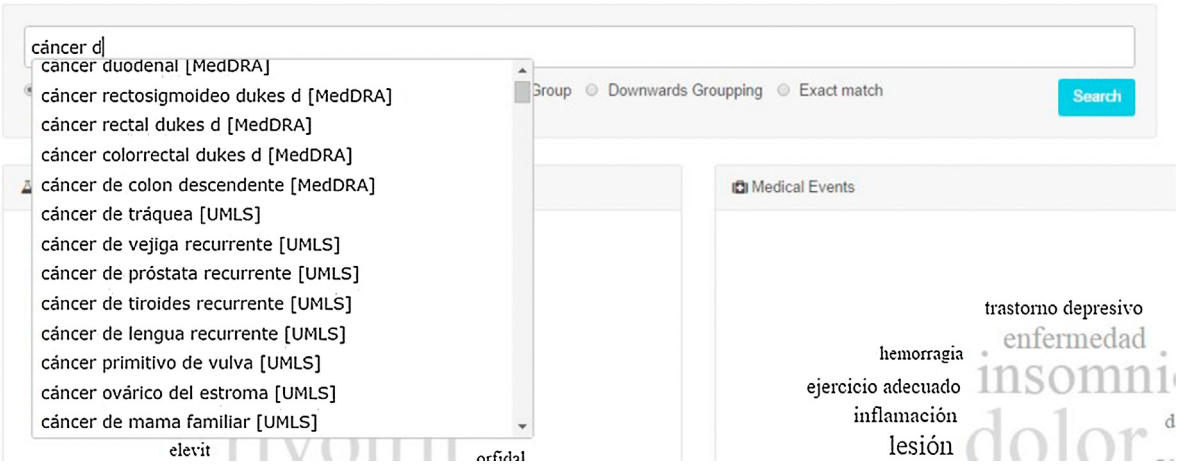

Fig. 7. Example showing search options using *cáncer* (cancer) query.

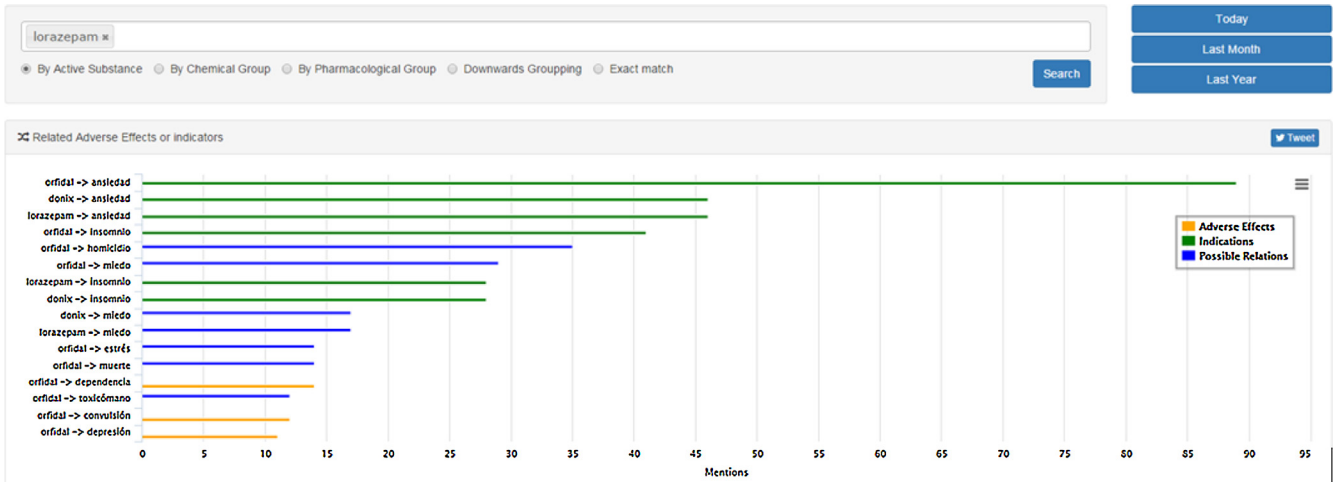

Fig. 8. Graph showing aggregated data about effects related to drug *Lorazepam* (indications, ADRs and possible relations).

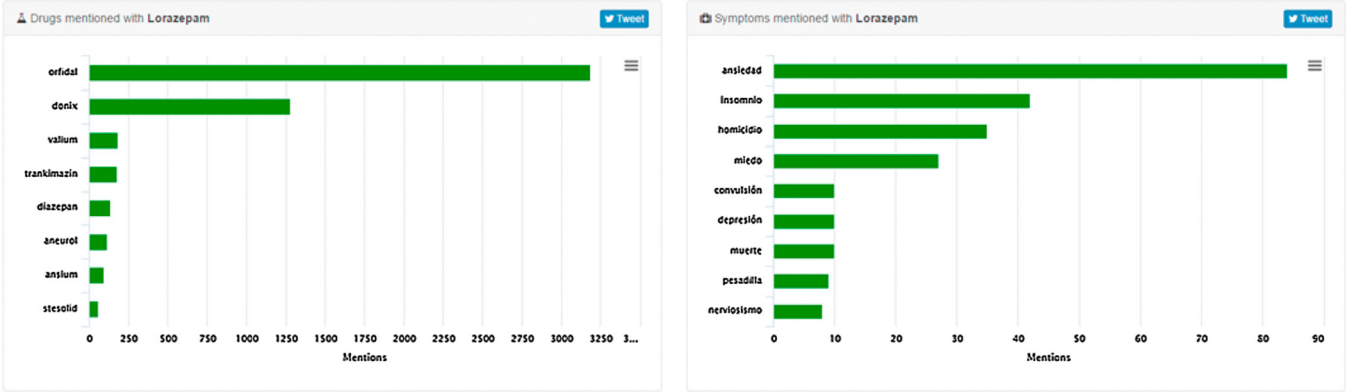

Fig. 9. Graph showing entity pairs aggregated data for *Lorazepam* active substance.

Finally, the system also presents information about the evolution of mentions through a timeline graph with different granularity (months, weeks and days) like the one shown in Fig. 10. All graphs have been developed using Highcharts [33]

4. Experiments evaluating NER and RE

In order to evaluate the linguistic processor, we have used a corpus extracted from ForumClínic[34], an interactive web page intended for patients to increase their degree of autonomy with

[33] http://www.highcharts.com/.

[34] http://www.forumclinic.org/.

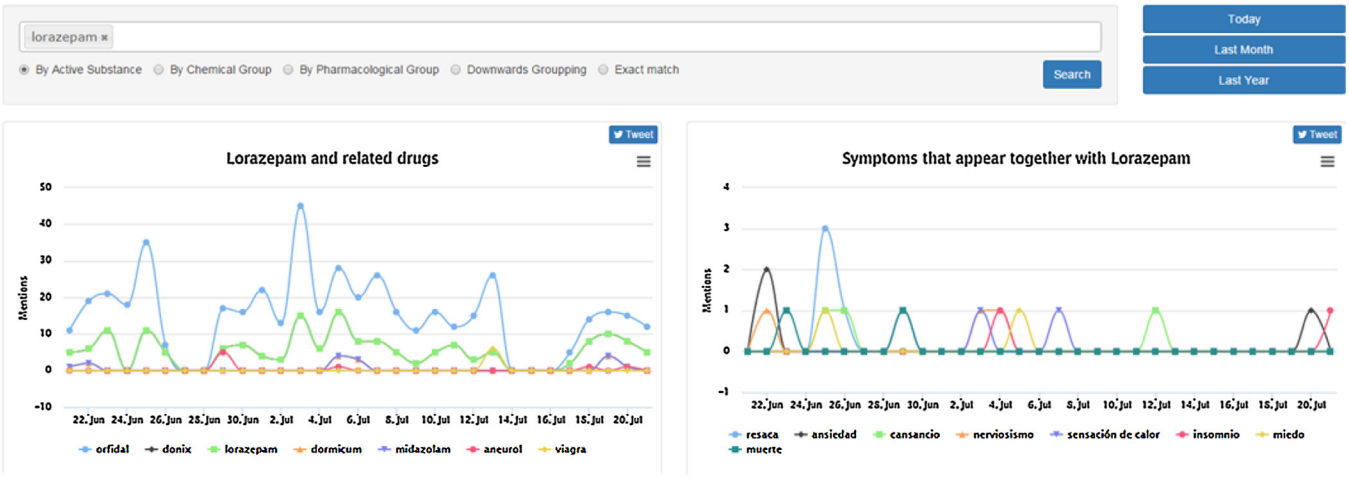

Fig. 10. Graph showing time-based evolution of entity mentions for *Lorazepam* query grouped by active substance.

respect to health issues, using the opportunities given by the newest Web technologies. Its target is to improve citizen's knowledge on health, diseases and their causes, as well as the efficiency and safety of the preventive treatments and medicines, so that they can get involved with the clinical decisions which attain them.

ForumClínic users are from all over the world, but a significant data is the fact that 46% of the webpage visits come from Spanish-speaking countries in America. In total, the number of a million users was reached in 2011, and it maintains a steady increase since it was created, in 2007. See Fig. 11 for information about structure and contents of this forum.

SpanishADR [26] is the first Spanish corpus annotated with drugs and effects by two annotators expert in the field; it consists of 400 user messages collected from ForumClínic. The size of the corpus is 26,519 tokens, whereas each message contains an average of 3.15 annotations (0.48 drugs, 1.42 effects and 1.25 relations). Moreover, it contains 189 drug annotations, 568 effect annotations and 164 drug–effect relations (the extension of SpanishADR corpus with drug-effect annotations is described in [24]). An assessment of the inter-annotator agreement (IAA) revealed that while drugs showed a high IAA (0.89), their effects pointed to moderate agreement (0.59). This may be due to drugs having specific names and being limited in number, while their effects are expressed by patients in many different ways due to the variability and richness of natural language.

We have evaluated the performance of the annotation pipeline described in Section 3.1 using the SpanishADR Gold standard. Metrics used are *precision* (*P*) and *recall* (*R*). Precision measures how many of the entities and relations that the system identified

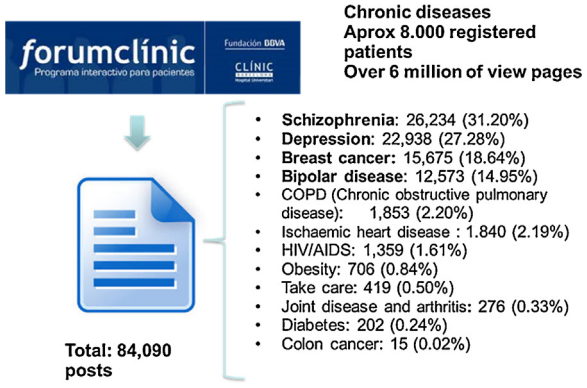

Fig. 11. ForumClínic structure.

were actually correct and recall represents how many of the entities and relations that should have been identified actually were identified. *F*-measure is the harmonic mean of precision and recall. P, R and *F*-measure are calculated according to two different criteria: the *strict* matching considers as correct every response where type entity and the spans are identical and the *lenient* matching considers every partially correct response as correct, i.e., the entity type is correct and the spans are overlapping but not identical.

A baseline has been defined using a system that integrates a gazetteer built from the CIMA resource in the case of drug recognition and a gazetteer created from the MedDRA database in the case of effects detection.

Regarding NER, Table 2 shows *P*, *R* and *F*-measure evaluating drug recognition and compared to the baseline defined. There is an increase of 0.36 in *F*-measure in strict matching mainly due to the fuzzy matched analysis done by the Topic recognizer. The main source of false negatives for drugs seems to be the abbreviations for drug families. For instance, *benzodiacepinas* (benzodiazepines) is commonly used as *benzos*, which is not included in our dictionary. An interesting source of errors to point out is the use of acronyms referring to a combination of two or more drugs. For instance, FEC is a combination of *Fluorouracil*, *Epirubicin* and *Cyclophosphamide*, three chemotherapy drugs used to treat breast cancer. Related to false positives some drug names such as *alcohol* (alcohol) or *oxígeno* (oxygen) can take a meaning different than the one of pharmaceutical substance. Another important cause of false positives is due to the use of drug family names as adjectives that specify an effect. This is the case of *sedante* (sedative) or *antidepresivo* (antidepressant), which can refer to a family of drugs, but also to the definition of an effect or disorder caused by a drug (sedative effects)

Table 3 shows *P*, *R* and *F*-measure evaluating effect recognition as well as the performance obtained with the baseline. In contrast to drug recognition, the difference with the baseline is smaller because ADRs are usually multiword terms and fuzzy matching is deactivated for them. The major source of false negatives was the use of colloquial and lay expressions to describe an effect. Patients

Table 2
Evaluation measures in drug recognition.

Drugs	NER system			Baseline		
	R	*P*	*F*-measure	*R*	*P*	*F*-measure
Strict	0.68	0.85	0.76	0.28	0.69	0.40
Lenient	0.68	0.85	0.76	0.32	0.77	0.43

Table 3
Evaluation measures in effect recognition.

Effects	NER system			Baseline		
	R	*P*	*F*-measure	*R*	*P*	*F*-measure
Strict	0.41	0.85	0.56	0.43	0.75	0.54
Lenient	0.47	0.83	0.60	0.42	0.86	0.56

used expressions such as *tengo la cabeza como un bombo* (my head is ringing) or *estoy destrozado* (I am destroyed) in order to express how they felt. These expressions are not included in our dictionary. A possible solution could be to create a lexicon containing these colloquial expressions. The second highest source of false negatives for effects was due to the different lexical variations of the same effect. For instance, *estrés* (stress) is a term included in our dictionary, but its lexical variations, like for example *estresado* (stressed), *estresante* (stressful), *me estreso* (I get stressed), *me estresa* (it makes me feel stressed) are not, and therefore they were not detected by our system. Nominalization may be used to identify all the possible lexical variations of a same effect. The third largest source of false negatives was spelling mistakes. We can see an example with *hurticaria*, which is an incorrect way of writing *urticaria* (urticaria). Many users have great difficulty in spelling unusual and complex technical terms. Nevertheless, although a fuzzy matching algorithm is used a more advanced matching method capable of dealing with more spelling error problems is required, mainly able to work with phrases. The last important error source was the use of abbreviations (*depre* is an abbreviation for depression), which also produces false negatives. Techniques such as lemmatization and stemming may help to cope with this kind of abbreviations.

False positives for adverse events were mainly due to the lack of ambiguity resolution. Some medical events receive the name of a common Spanish word, as it happens with *Zona* (Herpes zoster). Also acronyms used for long-named adverse events sometimes match with common words. For example, *Infección Respiratoria Alta*'s (Upper Respiratory Tract Infection) acronym, *IRA*, has different meanings in Spanish (past form of the verb to go or anger among others). Furthermore, some effects such as *anestesia* (anesthesia) share the name with the drug which drives patients to that state.

Table 4 shows *P*, *R* and *F*-measure evaluating relation extraction taking in to account drug–effect pairs annotated in the corpus (the objective is to evaluate relation extraction task regardless of NER task). Regarding the false positives, a cause of error is SpanishDrugEffectDB could include incorrect relations due to the fact that it was automatically obtained and it has not been manually revised. Another source of errors is the lack of context resolution. This means that, despite correctly detecting a drug and an effect (according to the drug package information), the context of the text did not fulfill the requirements to properly consider it a relation. Moreover, the lack of co-reference resolution introduces another important source of error for false positives; terms such as *enfermedad*, *efecto*, *tratamiento* and other have to be solved. An interesting source of errors is the lack of negation resolution, which means that despite the fact that the user specifies that he/she did not experience an effect after taking a drug, the system annotates the relation. Finally, the complex sentences (coordinated and subordinated sentences) in a comment may mislead the system into annotating a relation which is not correct.

Table 4
Evaluation measures in relation extraction (over drug–effect annotated pairs in Goldstandard corpus).

Window size		SpanishDrugEffectDB			Drug–effect pairs		
		R	*P*	*F*-measure	*R*	*P*	*F*-measure
30	Strict	0.08	0.57	0.14	0.63	0.44	0.52
	Lenient	0.13	0.96	0.24	0.88	0.61	0.72
100	Strict	0.10	0.34	0.16	0.74	0.26	0.38
	Lenient	0.23	0.74	0.35	0.99	0.34	0.51
250	Strict	0.12	0.32	0.17	0.17	0.75	0.33
	Lenient	0.24	0.67	0.36	1	0.29	0.45

Finally, concerning false negatives Table 4 shows that a great number of drug–effect pairs appearing in the corpus are not covered by the SpanishDrugEffectDB (recall is very low), that is, this database does not include all drugs effects. Therefore, the corpus has only 164 relations and it is difficult to conclude about the database coverage. Other studies are reported in [25] where a distant supervision method for relation extraction is analyzed using the overall ForumClínic corpus (84,000 user comments) as training and testing dataset.

5. Conclusions

In the healthcare scenario, there are three basic usages[35] of user-generated data that require special attention: (a) collecting information concerning behaviors of consumers by social media analytics, (b) diffusing messages and content to a wide audience via social media channels as an addition to other media such as web sites or news portals and (c) making people and organizations aware of healthcare issues leading to a public dialogue that could be viewed by anyone.

Focusing on the first one, many Internet health related sources (such as blogs, forums, etc.) and others such as EHR or clinical narratives contain a high volume of unstructured data mainly in form of free text. It is not only patient generated content but also clinician generated content. There is an increasing use of social networks (platforms to exchange the latest medical advances) by physicians [9].

This scenario poses several challenges and requires innovative ICT (Information and Communications Technology) products and services such as scalable NLP technology. Mining knowledge from health unstructured data has different applications: (a) Curating databases from free text sources (scientific articles, medical records, etc.). Adequate NLP tools can help curators by highlighting important sections of text (with relevant entities and relations) for review or even proposing an automatic interpretation along with an estimate of its accuracy. This prevents curators to analyze overall records and focusing on relevant content. (b) Monitoring specific healthcare items in social networks, for instance, pharmacovigilance and surveillance activities that currently are manually performed by domain experts, on-line monitoring patients' evolution or filtering contents to classify them. Medicines agencies are interested in monitoring the evolution of specific new drugs in the market, mainly drugs that could produce more adverse effects or black triangle drugs. (c) Codification tasks, related to entities annotation (drugs, diseases, symptoms among others) with specific vocabularies identifiers in order to transform pieces of text in structured knowledge, for instance, assign ICD codes (International Statistical Classification of Diseases and Related Health Problems) to diagnostics in clinical reports.

In this article, a system for monitoring medical events from social networks (drugs, diseases, symptoms and adverse effects) has been presented. It is based on a NLP pipeline built with MeaningCloud, a commercial Software as a Service platform providing customizable text analytics capabilities easy to integrate in any software application. The system is able to process real-time health related user generated content showing aggregated data about the different entities in several visualization timelines.

[35] Engaging patients through social media. Report by the IMS Institute for Healthcare Informatics, January 2014.

An evaluation has been also performed using a corpus annotated with drug–effect pairs [26] and an analysis of errors has been done with the aim of identifying future improvements. Drug NER performance is higher than ADR NER one. One issue that requires special attention is to manage patient oriented vocabulary. Patients do not report about their treatments using clinician terminology. Consumer Health Vocabulary[36] is a terminology for English language that contains lay terms but Spanish requires a similar resource that could be (semi)automatically built using NLP. Regarding the relation extraction task, as future work, we will first concentrate on improving the quality of the SpanishDrugEffectDB database. This database could be augmented from other websites about drugs and their effects in order to increase the recall of our system.

In addition, we will manually review the database in order to remove false positives, which are generated by the automatic process used to build the database. On the other hand, to improve the real-time performance of our system, we plan to apply text classification methods to automatically filter ADR related posts as is reported in [38].

Concerning NLP approaches, dictionary and rule-based technology has shown good performance in information extraction (IE) tasks. Chiticariu et al. [7] described a comparative of 54 rule-based products (including MeaningCloud) and concluded that only 1/3 of vendors rely on machine learning approaches. We believe that both approaches are complementary each other, for instance addressing ambiguity understanding texts. Therefore, such approaches must scale when they work with millions of records or tweets.

A fact that it is well known is that supervised learning methods currently achieve the best results for both NER and relation extractions tasks in any domain; but they require large, manually annotated training corpora. In this line, emerging trends are exploring the use of deep learning techniques applied to NER tasks that do not require large amounts of labeled data. Work described at Nikfarjam et al. [33] propose a sequence labeling method that uses word embedding features to mine ADRs mentions from user posts collected from DailyStrength and Twitter. They obtained better results compared to a lexicon-based baseline as well as a support vector machine (SVM)-based baseline. There are other emerging proposals such as [39], which explores the use of Twitter hashtags to extract hashtag-based networks that can help to expand the search space when querying drug-related literature. This means that social media can contribute with useful knowledge in linking and searching resources.

From a business point of view, every company should be aware of opinions and mentions about them given by their customers in Social Media as well as understanding their customers and businesses analyzing data in an adequate context to generate valuable knowledge. The healthcare agents should also be aware of this need. Health insurance companies and pharmaceutical companies are very interested in not only knowing when somebody talks about a brand or topic but also identifying if they are doing it on a positive or negative way. The value of such data is not entirely established mainly because mining and analysis of social media is an emerging science. Sentiment analysis of tweets or patient comments is a key aspect to detect consumer opinions about drugs that could be linked to adverse effects. For instance, if it is positive then the effect may be an indication or a beneficial effect; if it is negative then the effect may be a side effect.

Acknowledgments

This work was supported by TrendMiner project [FP7-ICT287863] and by eGovernAbility-Access project [TIN2014-52665-C2-2-R].

[36] http://www.consumerhealthvocab.org/.

References

[1] A.R. Aronson, F.M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (3) (2010) 229–236.
[2] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C.H. Leonarda, J.H. Holmes, Identifying potential adverse effects using the web: a new approach to medical hypothesis generation, J. Biomed. Inform. 44 (6) (2011) 989–996.
[3] Bian J., Topaloglu U., Yu F., Towards large-scale twitter mining for drug-related adverse events, in:, Proceedings of SHB, 2012, 25–32.
[4] C.A. Bond, C.L. Raehl, Adverse drug reactions in United States hospitals, Pharmacother.: J. Hum. Pharmacol. Drug Ther. 26 (5) (2006) 601–608.
[5] R.D. Boyce, P.B. Ryan, G.N. Norén, M.J. Schuemie, C. Reich, J. Duke, N.P. Tatonetti, G. Trifirò, R. Harpaz, J.M. Overhage, A.G. Hartzema, M. Khayter, E.A. Voss, C.G. Lambert, V. Huser, M. Dumontier, Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest, Drug Saf. 37 (8) (2014) 557–567.
[6] B. Chee, R. Berlin, B. Schatz, Measuring population health using personal health messages, in: AMIA Annual Symposium Proceedings 2009, American Medical Informatics Association, San Francisco, USA, 2009, pp. 92–96.
[7] Chiticariu, L., Li, Y., & Reiss, F.R., Rule-based information extraction is dead! Long live rule-based information extraction systems!, in:, Proceedings EMNLP 2013, (2013), pp. 827–832.
[8] P.M. Coloma, M.J. Schuemie, G. Trifirò, R. Gini, R. Herings, J. Hippisley-Cox, M. Sturkenboom, Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project, Pharmacoepidemiol. Drug Saf. 20 (1) (2011) 1–11.
[9] M.C. Domingo, Managing healthcare through social networks, Computer 43 (7) (2010) 20–25.
[10] C.C. Freifeld, J.S. Brownstein, C.M. Menone, W. Bao, R. Filice, T. Kass-Hout, N. Dasgupta, Digital drug safety surveillance: monitoring pharmaceutical products in twitter, Drug Saf. 37 (5) (2014) 343–350.
[11] Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., Smith, K., Gonzalez G., Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark, Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, BioTxtM 2014.
[12] H. Gurulingappa, A. Mateen-Rajput, L. Toldo, Extraction of potential adverse drug events from medical case reports, J. Biomed. Semant. 3 (1) (2012) 15.
[13] H. Gurulingappa, L. Toldo, A. Mateen-Rajput, J.A. Kors, A. Taweel, Y. Tayrouz, Automatic detection of adverse events to predict drug label changes using text and data mining techniques, Pharmacoepidemiol. Drug Saf. 22 (11) (2013) 1189–1194.
[14] R. Harpaz, W. DuMouchel, N.H. Shah, D. Madigan, P. Ryan, C. Friedman, Novel data-mining methodologies for adverse drug event discovery and analysis, Clin. Pharmacol. Ther. 91 (6) (2012) 1010–1021.
[15] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions, J. Biomed. Inform. 46 (5) (2013) 914–920.
[16] A. Herxheimer, M.R. Crombag, T.L. Alves, Direct Patient Reporting of Adverse Drug Reactions. A Twelve-Country Survey & Literature Review, Health Action International (HAI), Europe, 2010 (Paper Series Reference 01-2010/01).
[17] A. Kaplan, M. Haenlein, Users of the world, unite! The challenges and opportunities of social media, Bus. Horiz. 53 (1) (2010) 59–68.
[18] Leaman R., Wojtulewicz L., Sullivan R., Skariah A., Yang J., González G., Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks, in:, Proceedings of BioNLP 2010, 117–125.
[19] Q. Li, L. Deleger, T. Lingren, H. Zhai, M. Kaiser, L. Stoutenborough, A.G. Jegga, K.B. Cohen, I. Solti, Mining FDA drug labels for medical conditions, BMC Med. Inform. Decis. Mak. 13 (1) (2013) 53.
[20] M. McClellan, Drug safety reform at the FDA-pendulum swing or systematic improvement? N. Engl. J. Med. 356 (17) (2007) 1700–1702.
[21] Nikfarjam A., González G.H., Pattern mining for extraction of mentions of adverse drug reactions from user comments, in: Proceedings of AMIA Annual Symposium, 2011, 1019–1026.
[22] J. Parker, Y. Wei, A. Yates, O. Frieder, N. Goharian, A framework for detecting public health trends with Twitter, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, Niagara, ON, Canada, 2013, pp. 556–563.
[23] M. Rawlins, Pharmacovigilance: paradise lost, regained or postponed? The William Withering lecture 1994, J. R. Coll. Physicians Lond. 29 (1) (1995) 41–49.
[24] Segura-Bedmar I., Peña-González S., Martínez P., Extracting drug indications and adverse drug reactions from Spanish health social media, in:, Proceedings of BioNLP, 2014, 98–106.
[25] I. Segura-Bedmar, P. Martínez, R. Revert, J. Moreno-Schneider, Exploring Spanish Health Social Media for detecting drug effects, BMC Med. Inform. Decis. Mak. 15 (Suppl. 2) (2015) S6.
[26] I. Segura-Bedmar, R. Revert, P. Martínez, Detecting drugs and adverse events from Spanish social media streams, in: Proceedings of the 5th International Workshop on Health Document Text Mining and Information Analysis (Louhi), EACL, Gothenburg, Sweden, 2014.
[27] S. Sohn, J.P.A. Kocher, C.G. Chute, G.K. Savova, Drug side effect extraction from clinical narratives of psychiatry and psychology patients, J. Am. Med. Inform. Assoc. 18 (Suppl. 1) (2011) i144–i149.

[28] C.S. van Der Hooft, M. Sturkenboom, K. van Grootheest, H.J. Kingma, B. Stricker, Adverse drug reaction-related hospitalisations, Drug Saf. 29 (2) (2006) 161–168.
[29] WEB-RADR: new social media project for ADR monitoring in EU, 1519(1), 7–7, 2014.
[30] K. Wester, K.A. Jönsson, O. Spigset, H. Druid, S. Hägg, Incidence of fatal adverse drug reactions: a population based study, Br. J. Clin. Pharmacol. 65 (4) (2008) 573–579.
[31] R. Xu, Q. Wang, Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing, BMC Bioinform. 14 (1) (2013) 181.
[32] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, G. Gonzalez, Utilizing social media data for pharmacovigilance, J. Biomed. Inform. 54 (2015) 202–212 (C (April 2015)).
[33] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Am. Med. Inform. Assoc. 22 (3) (2015) 671–681.
[34] M. Neves, An analysis on the entity annotations in biological corpora, F1000Research 3 (2014) 96 (v1; ref status: indexed, http://f1000r.es/2o0).
[35] A. Neustein (Ed.), Text Mining of Web-Based Medical Content, De Gruyter, Berlin, Boston, 2014.
[36] L. Qian, G. Zhou, Tree kernel-based protein–protein interaction extraction from biomedical literature, J. Biomed. Inform. 45 (3 (June)) (2012) 535–543.
[37] I. Segura-Bedmar, P. Martínez, M. Herrero-Zazo, Lessons learnt from the DDIExtraction-2013 shared task, J. Biomed. Inform. 51 (October) (2014) 152–164 (ISSN 1532-0464).
[38] M. Yang, M. Kiang, W. Shang, Filtering big data from social media—building an early warning system for adverse drug reactions, J. Biomed. Inform. 54 (2015) 230–240 (C (April 2015)).
[39] A. Abdeen Hamed, X. Wu, R. Erickson, T. Fandy, Twitter K-H networks in action: advancing biomedical literature for drug search, J. Biomed. Inform. 56 (August) (2015) 157–168.
[40] I. Segura-Belmar, P. Martínez, Pharmacovigilance through the development of text mining and natural language processing techniques, J. Biomed. Inform. (2015) 58 (in press).

Paloma Martínez Paloma Martinez received the degree in Computer Science and the Ph.D. degree in Computer Science from the Universidad Politécnica de Madrid (Spain) in 1992 and 1998, respectively. Since 1992, she has been with the Advanced Databases Group in the Computer Science Department, Universidad Carlos III de Madrid. Her research lines are human language technologies (multilingual information extraction and retrieval in several domains, question answering, name entity recognition and relation extraction) as well as web accessibility. Some projects related to language technologies are TRENDMINER: Large-scale, Cross-lingual Trend Mining and Summarization of Real-time Media Streams (FP7-ICT 287863), ISSE: Semantic-based Interoperability for eHealth (FIT-350300-2007-75) and MULTIMEDICA project (Multilingual Information Extraction in Health domain and application to scientific and informative documents, TIN2010-20644-C03-01), devoted to research in technology based on automatic language processing for information extraction and retrieval from medical texts and other resources (reports, electronic medical records, social media, scientific documentation, etc.);. +info: labda.inf.uc3 m.es/pmf

José L. Martínez Business Development Manager at MeaningCloud LLC, holds a Ph.D. in Telecommunications at the Technical University of Madrid, and an Executive MBA from the IE Business School of Madrid. He has been working in the field of Information Access since he finished his grade, gathering experience in areas such as Information Retrieval, Information Extraction, Natural Language Processing, Business Intelligence, Semantic Technology and Big Data. During these years, he has taken part (first as an engineer and researcher, and then as a manager) in many projects involving information access technology to satisfy customer needs in different sectors (mass media, *telcos*, defense, energy, etc.). José Luis also enjoys teaching and, since 2002, he is a part-time professor at the Carlos III University of Madrid, in the Computer Science Department.

Isabel Segura-Bedmar Currently Visiting Associate Professor in Universidad Carlos III of Madrid. European Ph.D. in Computer Science, Universidad Carlos III de Madrid, April 2010. M.Sc. in Mathematics, Computer Science, Universidad Complutense de Madrid, 1998. I have been working on several research projects applying language technology to different problems and domains. My main area of research interest is the application of Information Extraction Techniques to the pharmacological domain, in particular, the extraction of drug–drug interactions and the detection of drug targets from biomedical texts. I obtained the Ph.D. degree with special distinction and award in 2010 and also holds the Award of the Spanish Society for the Natural Language Processing. I have been involved in the organization of several conferences and workshops (SEPLN 2008, BioSEPLN10), in addition to serving as a reviewer for the journal International Journal of Data Mining and Bioinformatics (IJDMB), Journal of Biomedical Informatics and committee program member of several workshops (DTMBIO10, HEALTHINF10).

Julián Moreno-Schneider Telecommunications Engineer since 2009 and Master in Computer Science and Artificial Intelligence from University Carlos III of Madrid. I am member of the Advanced Databases Group at Computer Science Department. My research interests include human language technologies (recovery and multilingual information retrieval in different domains and question answering), human-computer interaction and automation technology. In summer of 2012, I was doing an internship in DFKI (Saarbrücken, Germany) where I was collaborating in the ongoing projects reinforcing my knowledge by applying natural language processing techniques to the analysis of short news and tweets in the financial and political domain.

Adrián Luna Software Developer with experience in Data Science, Natural Language Processing and Web Development. Enjoy building new concepts into applications and always keen on exploring and experimenting with new technologies.

Ricardo Revert Telecommunication Engineer specialized in Telecommunications Planning and Management, by the Carlos III University of Madrid in 2014. In 2013, he joined the Advanced Databases Group (LaBDA) belonging to the Informatics Department of the Carlos III University, where he is currently working as a research staff member in the Natural Language Processing area related to Biomedicine.