UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA DE TELECOMUNICACIÓN

Detecting drugs and adverse events from Spanish health social media streams

Autor: Ricardo Revert Arenaz
Tutora: Isabel Segura-Bedmar

This page intentionally left blank

Título: DETECTING DRUGS AND ADVERSE EVENTS FROM SPANISH HEALTH SOCIAL MEDIA STREAMS

Autor: Ricardo Revert Arenaz

Tutora: Isabel Segura-Bedmar

EL TRIBUNAL

Presidente(a): _____

Vocal: _____

Secretario(a): _____

Realizado el acto de defensa y lectura del Proyecto fin de Carrera el día 30 de mayo de 2014 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO(A) PRESIDENTE(A)

Detecting drugs and adverse events from Spanish health social media streams

Ricardo Revert Arenaz

May 2014

Para ti, mamá

Abstract

Adverse Drug Reactions (ADRs) are the 4th cause of death in hospitalized patients. Despite the importance of clinical trials, they have many limitations mainly based on time and population. Therefore, other ways of spotting ADRs had to be created, as for instance, the healthcare professionals reporting systems and the spontaneous patients reporting systems created by the FDA or the EMA. Nevertheless, it has been proven that the results obtained are not yet as satisfactory as expected.

Health-related social media can be used along with these reporting systems in order to obtain possible information from a source where patients feel more comfortable sharing their experiences by exchanging information.

Therefore, the creation of the first corpus annotated with drugs and adverse events from social media in Spanish in order to train and evaluate machine-learning techniques is one of the main goals throughout this project. Furthermore, the implementation of a dictionary-based approach to detect mentions and to be tested with this gold-standard is the other main goal in this investigation.

Acknowledgments

Gracias Isa por volcarte en este proyecto y hacerlo mucho más fácil y divertido. Gracias Paloma por pensar desde un primer momento que podía hacerlo. Gracias Belén por obligarme a hacer el Bachillerato Tecnológico y por ayudarme como una segunda madre que eres para mi. Espero no haberos defraudado a ninguna.

Gracias Juli por malacostumbrarme a una vida que me ha gustado mucho. Gracias María por todos los ratos que hemos compartido en los últimos meses. Gracias Dani por ayudarme desde el primer día en toda duda que me surgía.

Gracias Junqui y Agui porque ya estamos aquí, y sabemos perfectamente que hubiera sido mucho más difícil si no lo hubiéramos hecho juntos y tan bien como lo hemos hecho.

Gracias Fouz por ser como un hermano en todo momento, por cada día de estudio y por todo el camino que hemos recorrido en paralelo que nos ha llevado a buen puerto.

Gracias a todos los amigos de toda la vida, que desde siempre han estado allí, y desde hace seis años comprendieron que esto es una ingeniería, y que iba a fallar más de lo que me hubiera gustado.

Tesekkurler bebeğim, porque viniste en un momento inmejorable para hacer de todo algo mejor y más bonito.

Y mamá, gracias por todo, especialmente por decirme que si no valía lo dejara, porque estoy aquí gracias a ti. No puedo decir que haya sido duro teniéndote como ejemplo en casa.

Contents

List of Tables

List of Tables

List of Figures

List of Figures

1. Introduction

1.1. Motivation

It is well-known that adverse drug reactions (ADRs) are an important health problem. Actually, ADRs are the 4th cause of death in hospitalized patients (Wester et al., 2008 [35]). Thus, the field of pharmacovigilance is currently receiving a great attention due to the high and growing importance of drug safety incidents (Bond and Raehl, 2006 [7]) as well as to their high associated costs (van Der Hooft et al., 2006 [29]).

Since many ADRs are not captured during clinical trials, the major medicine regulatory agencies such as the US Food and Drug Administration (FDA) or the European Medicines Agency (EMA) require healthcare professionals to report all suspected adverse drug reactions. However, studies show that ADRs are under-estimated due to the fact that they are reported by voluntary reporting systems (Bates et al., 2003 [4]; van Der Hooft et al., 2006 [29]; McClellan, 2007 [23]). In fact, it is estimated that only between 2 and 10 per cent of ADRs are reported (Rawlins, 1995 [31]). Healthcare professionals must perform many tasks during their workdays and thus finding the time to use these surveillance reporting systems is very difficult. Furthermore, healthcare professionals tend to report only those ADRs on which they have absolute certainty of their existence. Several medicines agencies have implemented spontaneous patient reporting systems in order for patients to report ADRs themselves. Some of these systems

are the MedWatch from the FDA, the Yellow Cards from the UK Medicines agency (MHRA) or the website[1] developed by the Spanish Agency of Medicines and Medical devices (AEMPS). Unlike reports from healthcare professionals, patient reports often provide more detailed and explicit information about ADRs (Herxheimer et al., 2010 [15]). Another important contribution of spontaneous patient reporting systems is to achieve patients having 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 are still unaware of their existence and even may feel embarrassed when describing their symptoms.

In this study, our hypothesis is that health-related social media can be used as a complementary data source to spontaneous reporting systems in order to detect unknown ADRs and thereby to increase drug safety. In recent days, social media on health information, just like has happened in other areas, have seen a tremendous growth (Hill et al., 2013 [16]). Examples of social media sites include blogs, online forums, social networking, and wikis, among many others. In this work, we focus on health forums where patients often exchange information about their personal medical experiences with other patients who suffer the same illness or receive similar treatment. Some patients may feel more comfortable sharing their medical experiences with each other rather than with their healthcare professionals. These forums contain a large number of comments describing patient experiences that would be a fertile source of data to detect unknown ADRs.

Although there have been several research efforts devoted to developing systems for extracting ADRs from social media, all studies have focused on social media in English, and none of them have addressed the extraction from Spanish social media. Moreover, the problem is that these studies have not been compared with each other, and hence it is very difficult to determine the current "state-of-art" of the techniques for ADRs extraction from social media. This comparison has not been performed due to the lack of a gold-standard corpus for ADRs.

[1] https://www.notificaram.es/

We hope our system will be beneficial to AEMPS as well as to the pharmaceutical industry in the improvement of their pharmacovigilance systems.

1.2. Objectives

Following with the motivation of this study, the main goal of our work is twofold: On the one hand, to create a gold-standard corpus annotated with drugs and adverse events and on the other hand, to develop a system to automatically extract mentions of drugs and adverse events from Spanish health-related social media sites. The corpus is formed by patients' comments from Forumclinic[2], a health online networking website in Spanish. This is the first corpus of patient comments annotated with drugs and adverse events in Spanish. Also, we believe that this corpus will facilitate comparison for future ADRs detection from Spanish social media.

This work has been performed in "TrendMiner: Large-scale Cross-lingual Trend Mining of Real-time media streams" project (FP7-ICT 287863), http://www.trendminer-project.eu/. LaBDA Research Group participates in this project. The goal of TrendMiner is to deliver innovative, portable open-source real-time methods for cross-lingual mining and summarisation of large-scale stream media. TrendMiner will achieve this through an inter-disciplinary approach, combining deep linguistic methods from text processing, knowledge-based reasoning from web science, machine learning, economics, and political science. Results will be validated in three high-profile case studies: financial decision support (with analysts, traders, regulators, and economists), political analysis and monitoring (with politicians, economists, and political journalists) and heatlh domain. The techniques will be generic with many business applications: business intelligence, customer relations management, community support. The project will also benefit society and ordinary citizens by enabling enhanced access to government data archives, summarisation of online health information, and tracking of hot societal issues.

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

1.2.1. Specific objectives

Below the specific objectives are shown:

- Create a technological study regarding the main developed systems to date which detect ADRs from social media.
- Carry out a study of the main health-based social media.
- Analyze the main terminological resources regarding drugs and ADRs in Spanish language.
- Study and analyze Textalytics, a multilingual tool intended for text analysis.
- Study and analyze the text analysis platform, GATE.
- Create a corpus annotated with drugs and adverse events which will be later used in order to develop and evaluate our extraction system.
- Calculate the inter-annotator agreement for the corpus. This metric will give us a clue on the degree of difficulty that the annotation task carries, and on the quality of the corpus' annotation.
- Build a dictionary for drugs and adverse events based on the aforementioned resources.
- Implement a system for the detection of drugs and adverse events by means of a dictionary-based approach.
- Evaluate our system with the corpus previously created.
- Conduct an error analysis in order to obtain the main causes of false positives and false negatives in our system.
- Analyze all possible approaches for future work which are intended to solve the errors detected in the past.

1.3. Roadmap

A brief overall view of the contents included in the document is next explained:

- Chapter 1 'Introduction': The brief introduction to the project, starting with the motivations which made us develop this project, continuing with

all the objectives to be covered and to ending up with an overview of the content in each chapter of the document.

- Chapter 2 'Background': The review of the social media analysis which took place at first, along with the analysis of the terminological resources both for drugs and adverse events, all of them in Spanish. Furthermore, the related work is included.
- Chapter 3 'The SpanishADRs corpus': The whole process which brought to life the first Spanish corpus annotated with drugs and adverse events. From its creation and annotation to the results which showed up.
- Chapter 4 'System description': The architecture and explanation of the dictionary-based approach that we implemented in order to extract drugs and adverse events from user comments from Spanish social media.
- Chapter 5 'Evaluation': The introduction to the way in which the system was evaluated, followed by the results obtained. The main focus is on the error analysis.
- Chapter 6 'Budget'. The planning and total budget of the project, including all the tasks related to the project, as well as the direct and indirect expenses.
- Chapter 7 'Conclusions and future work'. The learnings obtained throughout the realization of this project, as well as the future work which should be carried on in order to improve the performance.

2. Background

2.1. Information extraction applied to biomedical domain

The main investigation area in which to place this project is the Natural Language Processing (NLP). NLP manages ways to translate between computer and human languages, being its main objective to automate this translation process.

One step further, we focus on Information Extraction (IE), were we deal with automatically extracting structured information from unstructured machine-readable documents (in this case, social media). Basically, this task is based in processing human language texts by virtue of the aforementioned natural language processing.

The main motivation for these two fields came from competitions regarding recognition of named entities, such as people names or organizations from news articles.

This fact leads us to the next stop in our investigation: Named-Entity Recognition (NER). As we can figure out from the text above, NER is a sub-duty of the information extraction, with the goal of locating and classifying elements in a text into pre-defined categories.

Entities are generally noun phrases, normally of one to a few tokens length, which are found in unstructured text. The most common form of entity is the named entity (names of people, locations and companies). Nevertheless, nowadays the term entities has been expanded until including generics such as disease names, protein names, paper titles or journal names.

As a matter of fact, this investigation is based on the biomedical domain. More in concrete, in the Biomedical Named Entity Recognition (BNER), defined as the task of recognizing and categorizing entity names in biomedical domain. To date, most of the studies of BNER focused on genes and proteins, leaving the drug names more unattended. Some of the difficulties regarding BNER are:

- Drug names can be expressed not only by its name, but also with plurals, compounds and anaphoric expressions.
- Due to the fast changes in vocabulary, new drugs are constantly created, and older ones can be renamed. This makes it essential to have updated resources.
- Drug names are synonymous with other drug names.
- Acronyms and abbreviations are widely used in biomedical texts.

Focusing in this project, we will apply BNER techniques to recognize both drugs and adverse events from heath related social media, in Spanish language.

2.2. Review of the main social media in Spanish language

The first step towards the realization of this project was the decision of which social media was going to be used. In the first place, a large number of options were on top of the table (Twitter[3], Facebook[4], Tuenti[5], blogs, forums...), always under the condition of obtaining the contents in Spanish.

[3] http://twitter.com
[4] http://facebook.com
[5] http://tuenti.com

A detailed analysis of each and every of the options was then done separately in order to be able to decide the most appropriate media for the project's context, pharmacovigilance.

Finally, the decision of the social media from which to obtain the corpus to be used in the project was done.

Twitter

From the beginning Twitter was the first and most attractive option. The interest towards this social media is growing exponentially, as so is happening with the studies based on it. Companies are increasingly using it in order to detect market traces, understand their corporative reputation...

Twitter data speaks for itself, with one million new accounts created every day (11 per second), more than 300.000 daily visits and a total of approximately 500 million users (from which 140 million are active users).

Every second 750 tweets are sent in average, with a total of 175 million on a busy day. Another piece of relevant data is the number of searches done every month in Twitter, with a number of 24.000 million (against the 13.500 million combined between Yahoo! and Bing).

Never mind, on the other hand Twitter has also negative aspects. Firstly, regarding tweets, Twitter only allows access to 7 days of comments at one point in time. Therefore, the best option would be buying collections of tweets, which are available, and work with them.

Moreover, regarding pharmacovigilance, Graciela et al. (2013) [12] gathered tweets for one month, obtaining 42.327 of them which had a relationship to their set of drugs and diseases. From those, around half of them were advertisements, and from the useful ones (the rest of them), a small percentage was annotated and worked with. Only a 6.5% of this set included an adverse reaction. More specifically, removing the ones related to the term "nicotine", they had only between 1 - 2% left.

	COUNT
Total Twitter Data collected	42327
Tweets that are not advertisements	22739
Tweets manually annotated	3338
Tweets containing an adverse reaction	216

Table 1: Summary of retrieved tweets by Dr. Graciela González

Taking these results into account, and understanding that there is a very relevant additional handicap which is the language (Spanish), Twitter would have to be a discarded option to be the chosen social media for the corpus of this project.

Figure 1: Example of a tweet with an adverse event

Facebook

To this day, Facebook owes many more users than Twitter, being this number of around 1.250 million in total. Its main drawback when it comes to using this social media for this corpus is the fact that Facebook does not have such a well-defined timeline as Twitter, which makes the information extraction way more complicated.

On the other hand, the existence of Facebook Groups is really interesting, where users can exchange posts and messages related to a certain topic (for instance, related to a disease or drug, which would be useful for the corpus). Despite this, the access to these groups tends to be restricted to members, and the information in Spanish continues being limited in the context of pharmacovigilance. Therefore, Facebook is another discarded option.

Figure 2: Example of a Facebook comment with an adverse event

Blogs and Forums

Forums (with a 28% of active users) are the fourth most used social media in Spain (right after Facebook, YouTube[6] and Twitter). Blogs are just a few steps from forums, with a 17% of active users, as shown by The Cocktail Analysis[7]. The dynamic use of this couple of medias makes extracting information from them interesting enough.

A blog is a website which is periodically updated, and that contains texts or articles from one or several authors, being the owner the one who has the last word about what is then published under it.

Furthermore, a forum is a place for opinion and experience exchange, where everyone can write, creating new topics or adding some value to the existing ones.

Taking these two options into account, the higher dynamism and greater participation which characterizes forums makes them a more attractive choice for the project's context. Moreover, the existence of forums of all types and domain makes it easy to search for those which are related to the pharmacovigilance context.

In conclusion, due to all these advantages, forums are chosen as the social media to create the corpus. Among the forums of interest, the most important ones are:

[6] http://youtube.com
[7] http://tcanalysis.com/

- Forumclínic
- PortalesMédicos[8]
- EnFemenino[9]
- Saluspot[10]

Figure 3: Example of a forum post with an adverse event

Forum election

Once the decision was made to use forums as social media, an analysis of some of the options had to be done.

All forums follow the same structure. First, the creation of a main category (for instance, schizophrenia - *esquizofrenia*), under which people start new topics with the questions/comments/experiences they want to expose (for example, the adverse effects of a drug). Among these, people registered in those webs (or accessing as invited) leave their own comments, experiences, doubts, questions or answers, always related to the topic under which they are writing.

This makes it possible to obtain different experiences and points of view of a same topic. Regarding the project's context, this is something really positive when it comes to a medical or pharmaceutical issue.

The three aforementioned forums are quite similar between them, and therefore comments from all of them could be valid. Despite this, it has been decided to focus on one of them in particular. The chosen one has been Forumclinic, which will be deeply explained in section 3.1.

[8] http://www.portalesmedicos.com
[9] http://www.enfemenino.com
[10] https://www.saluspot.com/conocenos/quienes-somos-saluspot

2.3. Review of the main terminological resources for drugs and adverse events in Spanish language

Drug resources

The project required a complete, trustworthy and manageable drug resource. There are several of them on the Internet, with different levels of access difficulty.

The main barrier is, once again, the language, as the Spanish names of the drugs were needed. This means that some of the resources are directly discarded. After filtering, there were four main candidates:

- Snomed CT[11]
- Nomenclátor Digitalis - Integra[12]
- AEMPS's CIMA[13]
- Vademecum[14]

Snomed CT

Snomed CT (Systematized Nomenclature of Medicine – Clinical Terms) is a terminology whose content consists of concepts, descriptions and relationships, and has as a goal representing in a precise way their information and clinic knowledge.

It is owned, maintained and distributed by the International Health Terminology Standards Development Organisation[15] (IHTSDO), of which Spain is a member, making it accessible in Spanish.

[11] http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html
[12] http://www.msssi.gob.es/profesionales/farmacia/nomenclatorDI.htm
[13] http://agemed.es/cima/pestanias.do?metodo=accesoAplicacion
[14] http://www.vademecum.es
[15] http://www.ihtsdo.org/

In 2011, it included over 311.000 concepts with meaning and definition, organized into acyclic taxonomic hierarchies, and linked by around 1.360.000 links (the aforementioned relationships).

The Spanish Edition of the International Release is updated each year in April and October, which means that the last accessible version is the first one of 2013.

Once registered in the UMLS Terminology Services (UTS) in order to access the UMLS resources, a procedure which can take up to three days, you get access to the relational database which is organized in "concepts", "descriptions" and "relationships".

CONCEPTID	FULLYSPECIFIEDNAME	SNOMEDID	ISPRIMITIVE
329652003	Ibuprofen 200mg tablet (product)	C-50294	1
329653008	Ibuprofen 400mg tablet (product)	C-50295	1
329654002	Ibuprofen 600mg tablet (product)	C-50296	1

Table 2: Example of the drug "Ibuprofen(o)" in the Snomed CT's contents file

Despite having the Spanish translation in the aforementioned version, part of the contents are still in English, which does not suit the needs. Specifically, they are the drug names which are not translated, but have the English name (in the "concepts" file). Therefore, when searching for "Ibuprofeno", for instance, no result is achieved, as in the database it is shown as "Ibuprofen".

Due to this problem, this resource was discarded.

Nomenclátor Digitalis - Integra

Nomenclátor Digitalis-Integra is maintained, updated and distributed by the Spanish Ministry of Health (Ministerio de Sanidad, Servicios Sociales e Igualdad[16]), and therefore has national scope.

On the one hand, Nomenclátor Digitalis is a list containing relations between pharmaceutical products with its product identification, price and

[16] http://www.msssi.gob.es/

contribution to the user, used for medical prescriptions invoicing. On the other hand, Integra is a complementary file which contains different types of pharmaceutical products classified following the ATC system and including information about their content identification, DDD and administrative characteristics.

It contains 47.700 data entries (it includes the Digitalis DB resource), and it had its last update on March 2012.

It has a downloadable version (both in .txt and .mdb), with no need to register a website.

CODIGO	NOMBRE	NOMBRE2	CODATC	CAGR	CODDOE	CODDOH
984690	IBUPROFENO ROVI 500MG		M01AE01	0	P14793	H13375

Table 3: Part of an entry of the Integra document

Other resources were studied in case they gave a better performance.

AEMPS's CIMA

CIMA is a resource provided and maintained by The Spanish Agency for Medications and Healthcare Products (AEMPS[17] - Agencia Española de Medicamentos y Productos Sanitarios). It is an application which includes all authorized drugs in Spain. Its goal is to provide all information about each and all of them, and it allows the access to the drug's data sheet, as well as its patient information leaflet.

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 patient information leaflet is the document included inside the box of the medicine, and whose goal is to inform the patient.

The application encloses the following information related to the authorized drugs in Spain:

[17] http://www.aemps.gob.es/

- Drug's name
- Generic ingredient(s)'s name
- Marketing authorization holder's name
- National code
- Register number
- Authorization date
- Presentation name
- Presentation status (authorized / temporarily suspended / revoked)
- Availability in market
- Need and type of medical prescription
- Drug's data sheet
- Drug's package leaflet

All authorized drugs in Spain are codified, including the National Code, Fabrication Laboratory, Commercialization Laboratory, ATC Code or Generic Ingredient among other relevant information.

CIMA is a downloadable tool which consists in a relational database in xml format. In total, there are 15 documents, 14 of them including different and relevant information (see below) such as ATC codes, laboratories, or generic ingredients, and a main file (Prescripcion.xml), with the prescription of all the available drugs, including the information of the rest of the documents, codified in a proper and defined way.

```
<prescription>
  <cod_nacion>677471</cod_nacion>
  <nro_definitivo>70121</nro_definitivo>
  <des_nomco>AMOXICILINA/ACIDO CLAVULANICO ARDINECLAV 100/12,5 mg/ml POLVO PARA SUSPENSION ORAL EFG</des_nomco>
  <fecha_autorizacion>2008-10-02</fecha_autorizacion>
  <cod_sitreg>1</cod_sitreg>
  <fecha_situacion_registro>2008-10-02</fecha_situacion_registro>
  <fec_comer>2012-10-05</fec_comer>
  <des_prese>AMOXICILINA/ACIDO CLAVULANICO ARDINECLAV 100/12,5 mg/ml POLVO PARA SUSPENSION ORAL EFG, 1 frasco de 120 ml</des_prese>
  <cod_sitreg_presen>1</cod_sitreg_presen>
  <fec_sitreg_presen>2011-02-10</fec_sitreg_presen>
  <cod_vmpp>2941000140108</cod_vmpp>
  <cod_vmp>2911000140109</cod_vmp>
  <origen>AEMPS</origen>
  <oid_origen>1000140</oid_origen>
  <fecha_version>2011-12-26</fecha_version>
  <cod_atc>2471</cod_atc>
  <des_dosific>100 mg/ml + 12,5 mg/ml</des_dosific>
  <contenido>120</contenido>
  <unid_contenido>32</unid_contenido>
  <nro_conte>1 frasco de 120 ml</nro_conte>
  <cod_envase>6</cod_envase>
  <cod_forfar>9</cod_forfar>
  <cod_forfar_simplificada>65</cod_forfar_simplificada>
  <sw_receta>1</sw_receta>
  <sw_uso_hospitalario>0</sw_uso_hospitalario>
  <sw_diagnostico_hospitalario>0</sw_diagnostico_hospitalario>
  <sw_tld>0</sw_tld>
  <sw_especial_control_medico>0</sw_especial_control_medico>
  <sw_afecta_conduccion>0</sw_afecta_conduccion>
  <sw_psicotropo>0</sw_psicotropo>
  <sw_estupefaciente>0</sw_estupefaciente>
  <sw_tiene_excipientes_decl_obligatoria>1</sw_tiene_excipientes_decl_obligatoria>
  <sw_comercializado>1</sw_comercializado>
  <sw_presentacion_unidosis>0</sw_presentacion_unidosis>
  <sw_envase_clinico>0</sw_envase_clinico>
  <url_fictec>https://sinaem.agemed.es/DocumentosRAEFAR/2005/2005017780/HH_FT_001_001.doc</url_fictec>
  <url_prosp>https://sinaem.agemed.es/DocumentosRAEFAR/2005/2005017780/HH_PR_001_001.doc</url_prosp>
  <sw_sustituible>1</sw_sustituible>
  <sw_triangulo_negro>0</sw_triangulo_negro>
  <sw_foto_sensibilidad>0</sw_foto_sensibilidad>
  <sw_generico>1</sw_generico>
  <sw_huerfano>0</sw_huerfano>
  <sw_base_a_plantas>0</sw_base_a_plantas>
  <laboratorio_titular>2367</laboratorio_titular>
  <laboratorio_comercializador>2367</laboratorio_comercializador>
  <caducidad_ape>2ª Fase</caducidad_ape>
  <caducidad_pro>2ª Fase</caducidad_pro>
  <caducidad_rec>2ª Fase</caducidad_rec>
  <nro_pactiv>2</nro_pactiv>
  <viasadministracion>
    <cod_via_admin>54</cod_via_admin>
  </viasadministracion>
  <excipientes>
    <cod_excipiente>1801</cod_excipiente>
  </excipientes>
  <excipientes>
    <cod_excipiente>4885</cod_excipiente>
  </excipientes>
  <excipientes>
    <cod_excipiente>15025</cod_excipiente>
  </excipientes>
  <principiosactivos>
    <orden>1</orden>
    <cod_principio_activo>161</cod_principio_activo>
    <dosis>100</dosis>
    <cod_unidad_dosis>4</cod_unidad_dosis>
  </principiosactivos>
  <principiosactivos>
    <orden>2</orden>
    <cod_principio_activo>5419</cod_principio_activo>
    <dosis>12,5</dosis>
    <cod_unidad_dosis>4</cod_unidad_dosis>
  </principiosactivos>
</prescription>
```


Figure 4: Example of an entry from Prescripción

```
<atc>
  <nroatc>2471</nroatc>
  <codigoatc>J01CR02</codigoatc>
  <descatc>J01CR02 - Amoxicilina e inhibidores de la enzima</descatc>
</atc>
<atc>
```

Figure 5: Example of an entry from DICCIONARIO_ATC

```
<principiosactivos>
  <nroprincipioactivo>5419</nroprincipioactivo>
  <codigoprincipioactivo>419PK</codigoprincipioactivo>
  <principioactivo>CLAVULANATO POTASIO</principioactivo>
</principiosactivos>
<principiosactivos>
  <nroprincipioactivo>161</nroprincipioactivo>
  <codigoprincipioactivo>108TC</codigoprincipioactivo>
  <principioactivo>AMOXICILINA TRIHIDRATO</principioactivo>
</principiosactivos>
```

Figure 6: Example of two entries from DICCIONARIO_PRINCIPIOS_ACTIVOS

```
<laboratorios>
  <codigolaboratorio>2367</codigolaboratorio>
  <laboratorio>LABORATORIO REIG JOFRE, S.A.</laboratorio>
  <direccion>Gran Capita, 10.</direccion>
  <codigopostal>08970</codigopostal>
  <localidad>Sant Joan Despi (Barcelona)</localidad>
  <cif>A08259111</cif>
</laboratorios>
```

Figure 7: Example of an entry from DICCIONARIO_LABORATORIOS

Among all files, the following information can be obtained:

CONTENT	NUMBER	FILE
Drugs	16418	Prescripcion.xml
Generic ingredients	2228	DICCIONARIO_PRINCIPIOS_ACTIVOS.xml
ATC codes	1795	DICCIONARIO_ATC.xml
Laboratories	847	DICCIONARIO_LABORATORIOS.xml

Table 4: Summary of CIMA's content

Due to the high amount of information and the clarity of the content, CIMA was chosen as one of the two resources for drugs.

Vademecum

Vademecum is a guide of pharmaceutical products which is published and updated periodically, containing information provided by pharmaceutical companies.

In its online version, drugs, generic ingredients, laboratories, diseases, international equivalences of drugs, and drug interactions can be searched for.

Regarding drugs (over 18.200 drugs), they can be searched by name, clinical pharmacology, disease and symptoms, generic ingredient or laboratory. Generic ingredients can be accessed by name, clinical pharmacology and marketer laboratory.

Vademecum is the reference pharmacological guide in Spain, manageable and trustworthy, and therefore it was chosen to be part of the drug resource.

Adverse Event Resources

A total of 12 resources were analysed in order to come up with the best possible resource for adverse events. The 6 most important ones are:

- BOT Plus[18]
- MedLine Plus[19]
- Vademecum[20]
- MedDRA[21]

BOT Plus

The Health Knowledge Data Base, BOT Plus (Base de Datos del Conocimiento Sanitario, BOT Plus) is a computer program developed by the General Council of the Oficial Pharmaceutical Schools (Consejo General de Colegios Oficiales de Farmacéuticos) for the consult of homogeneous and updated information relative to drugs, health products, diseases and interactions.

[18] http://www.portalfarma.com/inicio/botplus20/que-es-Bot-Plus/Paginas/default.aspx#51
[19] http://www.nml.nih.gov/medlineplus/spanish/druginfo/meds/a681006-es.html
[20] http://www.vademecum.es/medicamentos-a_1
[21] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MDRSPA/termtypes.html

All the information regarding drugs and health products commercialized in Spain is retrieved from main official sources, such as AEMPS (Agencia Española de Medicamentos y Productor Sanitarios), EMEA[22] (European Medicines Agency) or pharmaceutical laboratories among others.

It contains all the updated information related to over 20.000 drugs and more than 2.000 generic ingredients commercialized in Spain.

Nevertheless, the difficult access to the resource makes it less attractive.

MedlinePlus

MedlinePlus is the National Institutes of Health's (NIH[23]) website intended for patients. The National Library of Medicine[24], which is part of the NIH, created and maintains MedlinePlus to give users authoritative up-to-date health information, such as diseases, conditions and wellness issues in a patient-oriented language.

It includes more than 900 health topics both in Spanish and in English, with information from over 1.000 organizations and more than 35.000 links to health information links.

Drugs and supplements can be browsed by generic or brand name, obtaining their information leaflets. Within the leaflet, a section for side effects of the medication is included, achieving known adverse effects of the drug in a patient-oriented language.

Again, the difficult access to the information makes the resource not interesting.

MedDRA

MedDRA (Medical Dictionary for Regulatory Activities) was developed by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH[25]) in the late 1990s. This

[22] http://www.ema.europa.eu/ema/
[23] http://www.nih.gov/
[24] http://www.nlm.nih.gov/
[25] http://www.ich.org/

medical terminology dictionary is clinically validated and used by the regulatory authorities in the pharmaceutical industry during the regulatory process.

The ICH MedDRA Management Board is the responsible for the direction of MedDRA, and is in charge of overseeing all the activities of the ICH MedDRA Maintenance and Support Services Organization (MSSO[26]), tasked to maintain, develop and distribute MedDRA.

MedDRA is a multilingual terminology allowing most users to operate in their native languages, as it has been translated and is maintained in the following languages: Chinese, Czech, Dutch, French, German, Hungarian, Italian, Japanese, Portuguese and Spanish. Thus, every term in MedDRA has an associated 8-digit code which is the same in all languages.

MedDRA has a very logical and well organised structure. It is composed of a five levels hierarchy, which goes from more very general to very specific. At the most specific level, called "Lowest Level Terms" (LLTs), there are 72072 terms which parallel how information is communicated. These LLTs reflect how an observation can be reported in practice.

On the next level, "Preferred Terms" (PTs) (20307 different ones), each member 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. Each LLT is linked to only one PT. Each PT has at least one LLT itself.

The "High Level Terms" (HLTs) group related PTs based on anatomy, pathology, physiology, etiology or function. In the same way, "High Level Group Terms" (HLGTs) groups HLTs.

The most general group, "System Organ Classes" (SOCs), group the HLGTs. They are groupings by etiology, manifestation site or purpose.

MedDRA is the adverse event classification dictionary approved by the ICH, and therefore a very reliable resource for the adverse events, thus it was chosen as one of the two resources for drugs.

[26] http://www.meddra.org/about-meddra/organisation/msso

Vademecum

This resource has been previously described as a drug resource.

Within the drug's information, the patient information leaflet is generally accessible. One of its paragraphs is dedicated to known adverse effects of the drug, written in a patient-oriented language.

This information is a valid support for the resource, more significantly taking into account that Vademecum was also chosen as a drug resource.

2.4. State of the art of drug-effect recognition in Social Media

In recent years, the application of Natural Language Processing (NLP) techniques to mine adverse reactions from texts has been explored with promising results, mainly in the context of drug labels (Gurulingappa et al., 2013 [14]; Li et al., 2013 [22]; Kuhn et al., 2010 [19]), biomedical literature (Xu and Wang, 2013 [34]), medical case reports (Gurulingappa et al., 2012 [13]) and health records (Friedman, 2009 [11]; Sohn et al., 2011 [32]). However, as it will be described below, the extraction of adverse reactions from social media has received much less attention.

In general, medical literature, such as scientific publications and drug labels, contains few grammatical and spelling mistakes. Another important advantage is that this type of texts can be easily linked to biomedical ontologies. Similarly, clinical records present specific medical terminology and can also be mapped to biomedical ontologies and resources. Meanwhile social media texts are markedly different from clinical records and scientific articles, and thereby the processing of social media texts poses additional challenges such as the management of meta-information included in the text (for example as tags in tweets) (Bouillot et al., 2013 [10]), the detection of typos and unconventional spelling, word shortenings (Neunedert et al, 2013 [25]; Moreira et al., 2013 [24]) and slang and emoticons (Balahur, 2013 [3]), among others. Moreover, these

texts are often very short and with an informal nature, making the processing task extremely challenging.

Regarding the identification of drug names in text, during the last four years there has been significant research efforts directed to encourage the development of systems for detecting these entities. Concretely, shared tasks such as DDIExtraction 2013 (Segura-Bedmar et al., 2013 [27]), CHEMDNER 2013 (Krallinger et al., 2013 [18]) or the i2b2 Medication Extraction challenge (Uzuner et al., 2010 [33]) have been held for the advancement of the state of the art in this problem. However, most of the work on recognizing drugs concerns either biomedical literature (for example, MedLine articles) or clinical records, thus leaving unexplored this task in social media streams.

Leaman et al., (2010) [20] developed a system to automatically recognize adverse effects in user comments. A corpus of 3,600 comments from the DailyStrength health-related social network was collected and manually annotated with a total of 1,866 drug conditions, including beneficial effects, adverse effects, indications and others. To identify the adverse effects in the user comments, a lexicon was compiled from the following resources: (1) the COSTART vocabulary (National Library of Medicine, 2008), (2) the SIDER database (Kuhn et al., 2010 [19]), (3) MedEffect[27] and (4) a list of colloquial phrases which were manually collected from the DailyStrength comments. The final lexicon consisted of 4,201 concepts (terms with the same CUI were grouped in the same concept). Finally, the terms in the lexicon were mapped against user comments to identify the adverse effects. In order to distinguish adverse effects from the other drug conditions (beneficial effects, indications and others), the systems used a list of verbs denoting indications (for example, help, work, prescribe). Drug name recognition was not necessary because the evaluation focused only on a set of four drugs: Carbamazepine, Olanzapine, Trazodone and Ziprasidone. The system achieved a good performance, with a precision of 78.3% and a recall of 69.9%.

An extension of this system was accomplished by Nikfarjam and Gonzalez (2011) [26]. The authors applied association rule mining to extract frequent patterns describing opinions about drugs. The rules were generated using the

[27] http://www.hc-sc.gc.ca/dhp-mps/medeff/index-eng.php

Apriori tool[28], an implementation of the Apriori algorithm (Agrawal and Srikant, 1994 [1]) for association rule mining. The system was evaluated using the same corpus created for their previous work (Leaman et al., 2010 [20]), and which has been described above. The system achieved a precision of 70.01% and a recall of 66.32%. The main advantage of this system is that it can be easily adapted for other domains and languages. Another important advantage of this approach over a dictionary based approach is that the system is able to detect terms not included in the dictionary.

Benton et al., (2011) [5] created a corpus of posts from several online forums about breast cancer, which later was used to extract potential adverse reactions from the most commonly used drugs to treat this disease: Tamoxifen, Anastrozole, Letrozole and Axemestane. The authors collected a lexicon of lay medical terms from websites and databases about drugs and adverse events. The lexicon was extended with the Consumer Health Vocabulary (CHV)[29], a vocabulary closer to the lay terms, which patients usually use to describe their medical experiences. Then, pairs of terms co-occurring within a window of 20 tokens were considered. The Fisher's exact test (Fisher, 1922 [9]) was used to calculate the probability that the two terms co-occurred independently by chance. To evaluate the system, the authors focused on the four drugs mentioned above, and then collected their adverse effects from their drug labels. Then, precision and recall were calculated by comparing the adverse effects from drug labels and the adverse effects obtained by the system. The system obtained an average precision of 77% and an average recall of 35.1% for all four drugs.

UDWarning (Wu et al., 2012 [36]) is an ongoing prototype whose main goal is to extract adverse drug reactions from Google discussions. A knowledge base of drugs and their adverse effects was created by integrating information from different resources such as SIDER, DailyMed[30], Drugs.com[31] and MedLinePlus. The authors hypothesized that unknown adverse drug effects would have a high volume of discussions over the time. Thus, the systems should monitor the number of relevant discussions for each adverse drug effect.

[28] http://www.borgelt.net/apriori.html
[29] http://consumerhealthvocab.org
[30] http://dailymed.nlm.nih.gov/dailymed/
[31] http://www.drugs.com/

However, to the best of our knowledge, the UDWarning's component devoted to the detection of unrecognized adverse drug effects has not been developed yet.

Bian et al., (2012) [6] developed a system to detect tweets describing adverse drug reactions. The systems used a SVM classifier trained on a corpus of tweets, which were manually labeled by two experts. MetaMap (Aronson and Lang, 2010 [2]) was used to analyze the tweets and to find the UMLS concepts present in the tweets. The system produced poor results, mainly because tweets are riddled with spelling and grammar mistakes. Moreover, MetaMap is not a suitable tool to analyze this type of texts since patients do not usually use medical terminology to describe their medical experiences.

As it was already mentioned, the recognition of drugs in social media texts has hardly been tackled and little research has been conducted to extract relationships between drugs and their side effects, since most systems were focused on a given and fixed set of drugs. Most systems for extracting ADRs follow a dictionary-based approach. The main drawback of these systems is that they fail to recognize terms which are not included in the dictionary. In addition, the dictionary-based approach is not able to handle the large number of spelling and grammar errors in social media texts. Moreover, the detection of ADRs has not been attempted for languages other than English. Indeed, automatic information extraction from Spanish-language social media in the field of health remains largely unexplored. Additionally, to the best of our knowledge, there is no corpus annotated with ADRs in social media texts available today.

3. The SpanishADR corpus

3.1. Corpus creation and annotation

One of the main goals of this project was to create the first corpus from social media in Spanish annotated with drugs and adverse events, and the first step was to obtain a first set of comments from the social media.

We must remind that, as it is specified in section 2.1, Forumclínic was the chosen forum to elaborate the corpus. Forumclínic is an interactive program intended for patients to increase their autonomy degree with respect to health, using the opportunities given by the newest technologies.

It provides rigorous, useful, transparent and objective information about health, whereas at the same time it boosts active participation of patients and associations.

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, being significant data the fact that 46% of the webpage visits come from Spanish America. In total, the million users were reached in 2011, and it maintains a steady increase since it was created, in 2007.

Corpus creation

The first step in order to obtain the corpus was implementing a web crawler which was able to gather all comments in Forumclinic. This was developed using Java. The total number of posts obtained from the site was of 84.090, distributed by topics in the following way:

- Esquizofrenia (*Schizophrenia*): 26.234 [31.20%]
- Depresión (*Depression*): 22.938 [27.28%]
- Cáncer de mama (*Breast cancer*): 15.675 [18.64%]
- Trastorno bipolar (*Bipolar disorder*): 12.573 [14.95%]
- EPOC (*COPD*): 1.853 [2.20%]
- Cardiopatía isquémica (*Coronary artery disease*): 1.840 [2.19%]
- VIH – Sida (*HIV*/*AIDS*): 1.359 [1.61%]
- Obesidad (*Obesity*): 706 [0.84%]
- Cuídate (*Take care*): 419 [0.50%]
- Artrosis y artritis (*Osteoarthritis and arthritis*): 276 [0.33%]
- Diabetes (*Diabetes*): 202 [0.24%]
- Cáncer de colon y recto (*Colon and rectal cancer*): 15 [0.02%]

All these comments were indexed with an alphanumeric code.

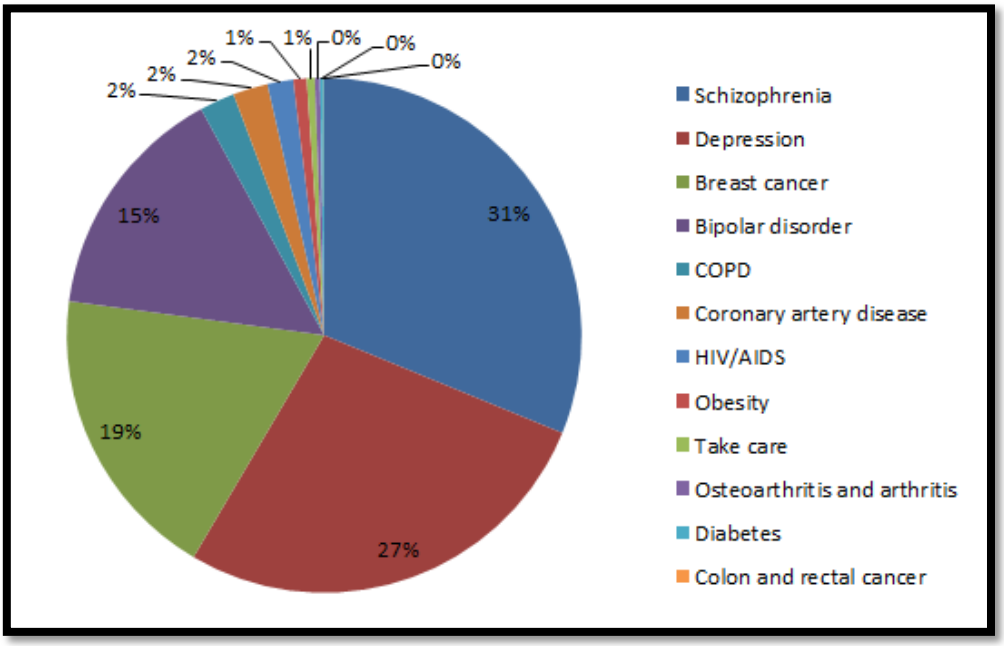

Figure 8: Forumclínic comments distribution by topic

A manually annotated corpus was then created consisting of 400 randomly selected posts (approximately 5% of the total). The quality and consistency of the annotation process was ensured through the creation of annotation guidelines.

Annotation Guidelines

An adverse event is considered to be negative, harmful. It includes any negative event that occurs to a patient while taking a drug, both if it is directly caused by it or not. For example: a patient takes part in a clinical trial with a drug. The responsible of this trial must follow the evolution of the patient, for both beneficial and adverse events, and note down all the experiences. This way, if the patient has "nausea", this should be annotated. The problem is that in that moment it is not possible to assure that "nausea" was caused by the drug or any other reason (for instance, the night before the patient could have eaten spoiled food). The responsible for the clinic trial would annotate "nausea" as adverse event.

A drug is a substance that is used in the treatment, cure, prevention or diagnosis of diseases. Several different types of drugs can be distinguished for the annotation, such as generic drugs, brand drugs or group of drugs.

DRUG TYPE	DESCRIPTION	EXAMPLE
Generic drug	Any chemical agent used in the treatment, cure, prevention or diagnosis of disease that has been approved for human use. It is denominated by a generic or chemical name, and not by a trade name or brand name.	Paracetamol
Brand drug	Any chemical agent used in the treatment, cure, prevention or diagnosis of disease that has been approved for human use. It is denominated by a trade or brand name.	Abilify
Family of drugs	A term in the text that describe different drugs into groups according to the organ or system on which they act or according to their chemical, pharmacological or therapeutic properties.	IMAO - Inhibidores de la Mono Amino Oxidasa (*MAOI – Mono Amine Oxidase Inhibitor*)

Table 5: Annotation guidelines referring to drugs

Taking into account these definitions, the decision was to annotate the corpus looking for terms or expressions which fulfilled one of them, organizing the tags in "Drug" and "Adverse Event".

Moreover, anaphoric expressions and orthographic errors regarding both sets are also annotated.

ANNOTATION	DESCRIPTION	EXAMPLE
Anaphoric expression	Presence of an element which makes reference to something previously mentioned, in this case, a drug or adverse event	...Yo tomaba Alprazolam. Es el ansiolítico que más dependencia genera... *(...I used to take Alprazolam. It is the anxiolytic which generates a higher dependency...)*
Orthographic error	An unintentional error of the accepted form of spelling a word, in this case, a drug or adverse event	Hemorrajia when referring to Hemorragia *(Hemorrhage)*

Table 6: Annotation guidelines referring to included terms

Corpus annotation

The sample of 400 comments was manually labeled with drugs and adverse events by two annotators with expertise in Pharmacovigilance. Disagreements between the annotators were discussed and reconciled during the harmonization process, where a third annotator helped to make the final. All the mentions of drugs and adverse events were annotated, even those containing spelling or grammatical errors. Nominal anaphoric expressions, which refer to previous adverse events or drugs in the comment, were also included in the annotation. The annotation tool used in the process was GATE[32].

The process was carried out following the annotation guidelines, which were created in an iterative process. The initial annotation guidelines included only generic and brand drugs, along with adverse events. This schema was discussed and developed by a team with a pharmacist, and two annotators (one of them a text mining expert), until the final guidelines were created.

[32] https://gate.ac.uk/

Disagreements

During the aforementioned harmonization process where the third annotator helped obtaining a final decision where the two first annotators did not coincide, cases like these occurred:

Orthographic error

De entre los distintos antiretrovirales, ***transcriptasa inversa****,* ***proteasa****,* ***integrasa*** *y* ***fusión****, qué grupo sería el más potente y cual el menos.*
(Among the different antiretroviral drugs, ***reverse transcriptase****,* ***proteinase****,* ***integrase*** *and* ***fusion****, which group would be the most powerful and which one the least)*

Table 7: Example of disagreement resulting in annotation

This case was consulted with a pharmacist due to the ambiguity. The terms in bold refer to a group of drugs (reverse transcriptase inhibitors, proteinase inhibitors, integrase inhibitors and fusion inhibitors). Nevertheless, they are not written in the correct way due to the absence of the term inhibitor, which gives the complete sense to the drug family. Therefore, it is considered an incorrect way of writing this family, and it will be annotated as a grammatical error of the form of a family drug.

Family of drug and anaphoric expression

Como ***complemento proteico*** *recomendamos el de los laboratorios Vegenat. Si compras los* ***complementos*** *del Decathlon, asegúrate que contenga* ***proteínas****.*
As a ***protein complement*** *we recommend the Vegenat laboratory's one. If you buy Decathlon's* ***complements****, make sure they contain* ***proteins****.*

Table 8: Example of disagreement resulting in non-annotation

All annotators agreed that *complement proteico* was a family of drugs. Furthermore, one annotator considered *complementos* is an anaphoric expression of this family, and *proteínas* as a drug ingredient, and therefore a drug. On the other hand, another annotator did not consider *complementos* or *proteínas* as relevant in the phrase. This is an example of the importance of an odd number of annotators, along with a pharmacist or expert in the field to obtain a consistent

and high quality annotated corpus. It worth mentioning that *complementos del Decathlon* was not annotated as a drug since it is not a brand-marked drug.

3.2. Inter-annotation agreement

To measure the inter-annotator agreement we used the F-measure metric. This metric approximates the Kappa coefficient (Cohen, 1960 [8]) when the number of true negatives (TN) is very large (Hripcsak and Rothschild, 2005 [17]). In our case, we can state that the number of TN is very high since TNs are all the terms that are not true positives, false positives nor false negatives. The F-measure was calculated by comparing the two corpora created by the two first annotators. The corpus labelled by the first annotator was considered the gold-standard. As it was expected, drugs exhibit a high IAA (0.89), while adverse events point to moderate agreement (0.59). As drugs have specific names and there are a limited number of them, it is possible to create a limited and controlled vocabulary to gather many of the existing drugs. On the other hand, patients can express their adverse events in many different ways due to the variability and richness of natural language.

The annotators found 187 drugs (from which 40 were nominal anaphors and 14 spelling errors) and 636 adverse events (from which 48 were nominal anaphors and 17 spelling errors). The corpus is available for academic purposes[33].

	DRUGS	ADVERSE EVENTS
Total annotations	187	636
Nominal anaphors	40	48
Spelling errors	14	7
IAA score	0.89	0.59

Table 9: Summary of the corpus annotation

[33] http://labda.inf.uc3m.es/SpanishADRCorpus

4. System description

4.1. Architecture

The architecture of the system is a simple pipeline in GATE, in which we embedded two plugins as modules. As we can see in **Figure 9: System architecture** one of them related to Textalytics[34] and the other one involving gazetteers.

Textalytics plugin is an entity dictionary which includes the drugs and the adverse events that are going to be marked by the system. The dictionaries include information from different resources: CIMA for the drugs and MedDRA for the adverse events.

On the other hand, the gazetteers (ANNIE Gazetteers) are lists we included in order to improve the performance of the system and to have a more complete set of resources. They include information from the WHO ATC system[35] and from the site Vademecum, both to expand the information on drugs and on adverse events.

[34] http://www.textalytics.com
[35] http://www.whocc.no/atc_ddd_index/

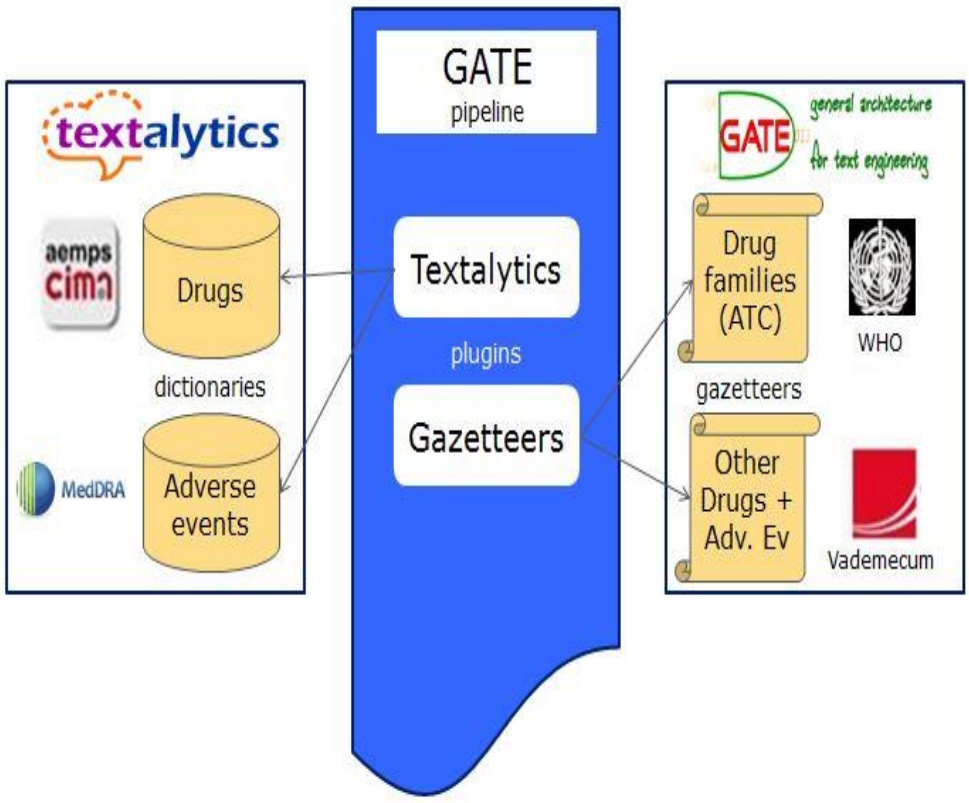

Figure 9: System architecture

GATE

GATE is an infrastructure for developing and deploying software components that process human language. It is open source free software which is in active use for any kind of computational task involving human language. Thus, it allowed us to create our system with the aforementioned pipeline architecture in order to process the posts from the forum.

As an architecture, GATE suggests that the elements of software systems that process natural language and successfully be broken down into several types of component, known as resources. In our case, these resources are the dictionary and the gazetteers that we mounted on the system, full of biomedical terminology oriented to the goal of detecting drugs and adverse events in social media.

When using GATE to develop language processing functionalities for an application, the developer uses GATE Developer and GATE Embedded in order to create resources, what involves programming and developing language resources.

Once the appropriate set of resources is created, they can be embedded in the system using GATE Embedded.

Textalytics

Daedalus[36] is a Spanish company specialized in automatically extracting the meaning of all kind of multimedia contents, by applying semantic technologies, language processing, voice recognition and data and text analysis.

Textalytics is a multilingual text analysis engine to extract information from any type of texts such as tweets, posts, comments, news, contracts, etc. This tool offers a wide variety of functionalities such as text classification, entity recognition, concept extraction, relation extraction and sentiment analysis, among others. We used a plugin that integrates Textalytics with GATE. In this project, we applied entity recognition provided by Textalytics, which follows a dictionary-based approach to identify entities in texts. We created a dictionary for drugs and adverse events from CIMA and MedDRA. This dictionary was integrated into Textalytics.

4.2. Construction of a dictionary for drugs and adverse events

Since our goal is to identify drugs and adverse events from user comments, the first challenge is to create a dictionary that contains all of the drugs and known adverse events.

4.2.1. Textalytics' dictionary

Textalytics is one of the analysis resources which will allow the automatic annotation of the corpus in our system. The Topic Extraction (v1.1) API is the solution for extracting the drugs and adverse events present in the corpus. This process is carried out combining a number of natural language processing techniques which allow obtaining morphological, syntactic and semantic analyses of the corpus in order to identify the aforementioned entities.

[36] http://www.daedalus.es

In order to identify drugs and adverse events, a user dictionary has to be created. The dictionary has the following structure:

```
' id \ form \ alias{alias1|aliasN} \ pos_tag \ lemma \ checkinfo_tag \ sense \
other_info \r \n '
```

Figure 10: Textalytics dictionary structure

The id is an aleatory number to identify the dictionary's entry, the form is the entry, and the alias will be those entities which will be annotated and related to the entry. The rest of the information is not important in this section.

4.2.2. Drug dictionary

Despite choosing two resources for drugs, only CIMA was used for creating the Textalytics drug dictionary. This decision was taken, as for the adverse events, due to the organized and relational structure of the resource.

After studying the files in CIMA, the decision was to create a dictionary in which the entries were the generic ingredients, and the aliases were those drugs which contained each generic ingredient. With this schema, all drugs and ingredients would be available for annotation, and the drug part of the dictionary would not be too large. As a matter of fact, this part consisted of 2228 entries (one per generic ingredient in CIMA), and a total of 3662 brand names as aliases of the first ones, being so when containing the generic ingredient.

In order to accomplish this objective, the two files to work with were Prescripcion.xml and DICCIONARIO_PRINCIPIOS_ACTIVOS.xml. The basic needs were obtaining the relations between drugs and their generic ingredients, and the brand name for the long and tedious drug names.

Drugs and generic ingredients relation

Prescripcion.xml is the main file in the CIMA resource. It is the one which covers all information related to all drugs (16418) included in it. This information is codified (ATC code appears with “cod_atc”, generic ingredient is shown with “cod_principio_activo”, laboratories as “laboratorio_titular” and “laboratorio_comercializador”...), and therefore the information has to be decoded.

For example, for the drug with “National Code” 600000, which receives the name of AMOXICILINA/ACIDO CLAVULANICO SALA 500/50 mg POLVO PARA SOLUCION INYECTABLE Y PARA PERFUSION EFG, 100 viales, in the Prescripcion.xml file, some of the important elements, including the generic ingredient, are shown as follows:

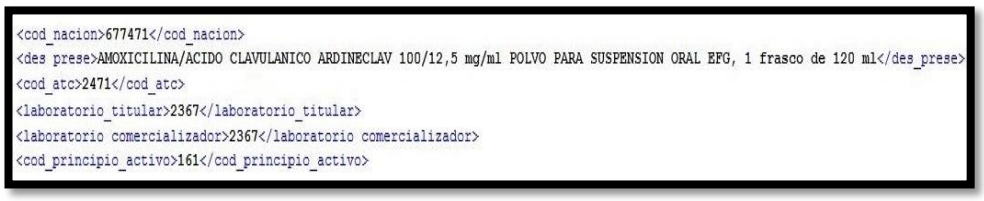

Figure 11: Extract with relevant information from a drug in Prescripcion

With this structure, other files are needed in order to obtain the desired information. The file DICCIONARIO_PRINCIPIOS_ACTIVOS.xml includes the information needed to decode and relate the generic ingredients with the drugs where they are included:

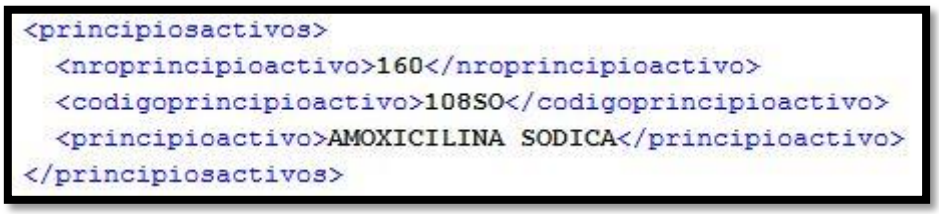

Figure 12: Entry of a generic ingredient in DICCIONARIO_PRINCIPIOS_ACTIVOS

The file DICCIONARIO_PRINCIPIOS_ACTIVOS.xml contains a total of 2228 generic ingredients registered in CIMA. All the drugs in Prescripcion.xml contain one or more of these active ingredients.

Following the above schema, all drugs in CIMA were related to their generic ingredient code. The ingredients will be the entries of the dictionary, and the brand names (will now be explained) the aliases.

Brand name

The Prescripcion.xml file contains a total of 16418 instances that are all brand drugs registered in CIMA. A brand drug is a medication marketed by a pharmaceutical company under a trademark name. The aim of the brand name is to give a short name for each brand drug. In principle, brand names hold more probability to be named or written by the users of social media.

The brand drugs saved in the Prescripcion.xml have usually very long names containing the brand name as well as information about their dosages, their form and administration method (for example, TAMOXIFENO LEPORI 10 mg COMPRIMIDOS). It is highly unlikely that these long names are used by patients. Therefore, the goal is to obtain their short names (without containing information on the dosage, form or route of administration).

In order to compute the brand name, the starting point is the complete brand drug name, which is compiled from the "des_nomco" element (which is child of the prescription element), in the Prescripcion.xml.

In order to do a differentiation between names, it is important to introduce the concept of Pharmaceutical Generic Specialty (in Spanish, EFG - Especialidad Farmacéutica Genérica). When a drug is an EFG, its name will be as follows:

Generic Ingredient + Laboratory + Dose/Pharmaceutical Form Info + EFG (+ Content Info)

Figure 13: EFG drug name structure

An example from CIMA would be:

PARACETAMOL KABI 10 mg/ml SOLUCION PARA PERFUSION EFG, 10 viales de 100 ml

Figure 14: CIMA example of EFG drug

On the other hand, drugs which are not EFG do not follow a strict pattern. Anyway, they can be considered as brand name drugs, where normally the generic ingredient does not appear in the name, whereas a commercial name does. In general, the structure would be:

Brand name + Dose/Pharmaceutical Form Info + Content Info

Figure 15: Non-generic drug name structure

An example from CIMA would be:

VALCYTE 450 mg COMPRIMIDOS RECUBIERTOS CON PELICULA., 60 comprimidos

Figure 16: CIMA example of non-generic drug

From this moment on, they will be considered as EFG drugs, and brand drugs.

Obtaining the brand name

First of all, a preprocessing is done to all the drug's long names. In this first step, some inconsistencies, redundancy and irrelevant information (such as Content Info) from the long names are removed.

1. Moving on, the next step is looking for a digit in the name. In general, all of the drugs in the resource contain a number, which is referring to their dosages (‘10’ for the EFG drug and ‘450’ for the brand drug in the examples) followed by a dose unit (‘mg/ml’ for EFG, ‘mg’ for brand). If the drug is filtered with this constraint, the result would be, for the EFG drugs, the Generic Ingredient + Laboratory, and for the brand drugs, directly the commercial name, by which it is known.
2. For each EFG drug, the laboratory name should also be removed in order to obtain only the name of its generic ingredient. Therefore, the next step will be detecting the laboratory name and removing it in order to achieve the active ingredient by itself. It is important to remember that this affects only to the EFG drugs.

   With the CIMA documentation it is possible to obtain a list of laboratories which will be very useful for this purpose. To do so, in the DICCIONARIO_LABORATORIOS.xml, as a child of the laboratorios element, the "laboratorio" elements contain the registered name of the laboratories registered in CIMA.

```
<laboratorios>
  <codigolaboratorio>2335</codigolaboratorio>
  <laboratorio>LABORATORIOS RAMON SALA, S.L.</laboratorio>
  <direccion>Paris 174-174 bis</direccion>
  <codigopostal>08036</codigopostal>
  <localidad>Barcelona</localidad>
</laboratorios>
```

Figure 17: Entry of a laboratory in DICCIONARIO_LABORATORIOS

The main problem is that the complete registered name of the laboratory is not the one mentioned in the long name of the drug. For instance, in the previous example, the full name is “PARACETAMOL KABI 10 mg/ml

SOLUCION PARA PERFUSION EFG, 10 viales de 100 ml". Following the rules, the laboratory would be "KABI", but the registered name of this laboratory is "FRESENIUS KABI ESPAÑA, S.A.U.". Thus, in order to obtain a Laboratory List with the laboratory names which are mentioned in the drug names, a filtering process has to be done to the long laboratory names (similar to the one done to the long brand names to obtain the short brand names).

Based on the observation of the list of long laboratory names, a list of stop words was compiled, including terms which can be removed from the registered laboratory names in order to obtain their equivalents as they are found in the drug name. Some examples of the most common stop words and their meaning (acronyms and foreign terms) are shown in the table below.

STOP WORD FOR LABORATORY	MEANING
LABORATORIO	
FARMACÉUTICA	
SPAIN	
ESPAÑA	
UK	
DEUTSCHLAND	
ARZNEIMITTEL	Drug in German
LIMITED	S.L. in English
LTD	LIMITED abbreviation
S.A.	[acronym] Sociedad Anónima
S.L.	[acronym] Sociedad Limitada
S.R.L.	[acronym] Sociedad de Responsabilidad Limitada
GMBH	[acronym] Gesellschaft Mit Beschränkter Haftung (S.R.L. in German)
AG	[acronym] Aktiengesellschaft (S.A. in German)
A/S	[acronym] Aktieselskab (S.A. in Danish)
AB	[acronym] Aktiebolag (S.L. in Swedish)
B.V.	[acronym] Besloten Vennootschap (S.R.L. in Dutch)

Table 10: Most frequent stop words for laboratories

3. Unfortunately, not all drug names follow the previous patterns (EFG or brand). An example would be:

Figure 18: CIMA example of a drug name with no pattern

The only identifiable part of the name is the Content Info. Apart from that, there is not a clear generic ingredient (the term *AGUA* appears in four generic ingredients, but not by itself or accompanied by *BIDESTILADA*), laboratory name or detectable brand.

These cases will be treated in a particular way, filtered with a list of stop words. The list will be compiled with words from different lists in CIMA, and a set of synonyms and related words.

- The first words are found in CIMA's file DICCIONARIO_UNIDAD_ADMINISTRACION, DICCIONARIO_FORMA_FARMACEUTICA [.xml]. In the first one, under the elements "unidadadministracion", child of unidadesadministracion, there are terms related to drugs administration, such as *COMPRIMIDOS*, *JERINGA* or *PASTILLAS*. Moreover, in the second one, in "formafarmaceutica" tag, child of the also called formafarmaceutica element, terms describing pharmaceutical forms are found, as for example, *SOLUCIÓN ORAL*, *POLVO PARA INHALACIÓN* or *POMADA NASAL*. Within these terms, stop words are retrieved (*SOLUCIÓN*, *POLVO*, *POMADA*...). They will be a first approach (around 70 words) to the final list of stop words which will be used to filter the drugs.
  For instance, regarding the example above, the stop word *AMPOLLAS* is detected, and the brand name for the drug will result as *AGUA BIDESTILADA.*
- Despite the use of this first list helped cleaning the names, there were still some words similar to the ones obtained from the CIMA files, but which did not appear in these. This leads to the second kind of words. For example, the term *PASTILLA* (i.e. MUCOSAN 15 mg PASTILLAS DE GOMA, 40 pastillas) was part of the Administration Unit List (compiled from DICCIONARIO_UNIDAD_ADMINISTRACION.xml), and therefore was

defined as a stop word; but on the other hand, a similar word like *CARAMELO* (i.e. CHICLIDA CARAMELO, 6 pastillas) was not in the CIMA document, thus it does not appear as a stop word. They both refer to administration units of a drug, so they both should be on the list. These words were detected by observation (going over the resulting drug names and spotting them) and added to the stop word list. More examples of these words would be *TABLETA*, *REPETABS*, *COLUTORIO* or *FILM,* all of them pharmaceutical forms but which did not appear in CIMA.

- One step forward was including in the list of stop words those which did not contribute enough in the differentiation of two drugs. For instance, ADOLONTA 100 mg/ml GOTAS ORALES EN SOLUCION, 1 frasco de 30 ml is the long name for a drug with National Code 665364. Moreover, the fabrication and commercialization laboratories are both GRÜNENTHAL PHARMA, S.A., its ATC code is N02AX02 - Tramadol, and it consists of one generic ingredient, TRAMADOL HIDROCLORURO.

  On the other hand, there is another drug, ADOLONTA RETARD 200 mg COMPRIMIDOS DE LIBERACION PROLONGADA, 20 comprimidos whose National Code is 665588. In this case, the laboratory is again GRÜNENTHAL PHARMA, S.A. both for fabrication and commercialization, the ATC code is the same, N02AX02 - Tramadol, and also the generic ingredient, TRAMADOL HIDROCLORURO.

  This is something quite common among drugs. Their composition can be varied by just some different excipients, which are generally inactive substances which are mixed with the generic ingredients (which in these cases would be the same) in order to give consistency and pharmacological characteristics to a drug.

  In the previous example it was appreciated how a very similar drug is differentiated in its name by a possible key word (or stop word) such as “RETARD”. Due to the previous explanation, these two drugs could share the same brand name (ADOLONTA).

  As it happens with this example, this is something quite common along the drugs in CIMA, and therefore there is a great number of stop words of

this kind. Examples of them would be *NIÑOS*, *ADULTOS*, *INFANTIL*, *INFANT*, *JUNIOR*...

- Still, there are some words which do not add any value to the brand name, and therefore should not be part of it. These are, for example, *INICIO*, *CENTRAL*, *PERIFERICO* or *CHOQUE* among others

Summing up, the total number of words in the list of stop words was of 235, some of them shown in the following table:

STOP WORD	**EXAMPLE IN CIMA**
NASAL	BACTROBAN 20 mg/g POMADA BACTROBAN *NASAL*, POMADA
RECTAL	ABRASONE CREMA ABRASONE *RECTAL*
PLUS	ACEOTO SOLUCION ACEOTO *PLUS* 3/0,25 mg/ml GOTAS ÓTICAS EN SOLUCIÓN
NIÑOS	APIRETAL 100 mg/ml SOLUCION ORAL APIRETAL *NIÑOS* 250mg SUPOSITORIOS
ADULTOS	FEBRECTAL 650 mg COMPRIMIDOS FEBRECTAL *ADULTOS* 600 mg SUPOSITORIOS
INFANTIL	FLUMIL 200 mg GRANULADO PARA SOLUCION ORAL FLUMIL *INFANTIL* 100 mg GRANULADO PARA SOLUCION ORAL
INFANT	AMINOVEN 10% SOLUCION PARA PERFUSION AMINOVEN *INFANT* 10% SOLUCION PARA PERFUSION
JUNIOR	FRENADOL COMPRIMIDOS EFERVESCENTES FRENADOL *JUNIOR*
UNIDIA	CLARITROMICINA TEVA 250 mg COMPRIMIDOS RECUBIERTOS CON PELICULA EFG CLARITROMICINA *UNIDIA* TEVA 500 MG COMPRIMIDOS DE LIBERACION PROLONGADA EFG
	RISEDRONATO TEVA 75 mg COMPRIMIDOS RECUBIERTOS CON PELICULA EFG

SEMANAL	RISEDRONATO *SEMANAL* SANDOZ 35 mg COMPRIMIDOS RECUBIERTOS CON PELICULA EFG
DIARIO	DECAPEPTYL *DIARIO* 0,1 mg, POLVO Y DISOLVENTE PARA SOLUCION INYECTABLE
MENSUAL	DECAPEPTYL *MENSUAL* 3,75 mg, POLVO Y DISOLVENTE PARA SUSPENSION INYECTABLE
TRIMESTRAL	DECAPEPTYL *TRIMESTRAL* 11,25 mg POLVO Y DISOLVENTE PARA SUSPENSION INYECTABLE
SEMESTRAL	DECAPEPTYL *SEMESTRAL* 22,5 mg POLVO Y DISOLVENTE PARA SUSPENSION INYECTABLE
RETARD	ADOLONTA 100 mg/ml GOTAS ORALES EN SOLUCION ADOLONTA *RETARD* 200 mg COMPRIMIDOS DE LIBERACION PROLONGADA
FORTE	AERO-RED 120 mg COMPRIMIDOS MASTICABLES AERO-RED *FORTE* 240 MG CAPSULAS BLANDAS
COMPLEX INSTANT	BEKUNIS TISANA BEKUNIS *COMPLEX* COMPRIMIDOS RECUBIERTOS BEKUNIS *INSTANT*, POLVO PARA SOLUCIÓN ORAL
FLAS	DESLORATADINA COMBIX 5 mg COMPRIMIDOS RECUBIERTOS CON PELÍCULA EFG DESLORATADINA *FLAS* COMBIX 5 mg COMPRIMIDOS BUCODISPERSABLES EFG
EMULGEL	VOLTAREN 50 mg COMPRIMIDOS GASTRORRESISTENTES VOLTAREN *EMULGEL* 1%
GEL SPRAY	DALGEN *GEL* DALGEN *SPRAY* SOLUCION

Table 11: Most frequent stop words for drugs

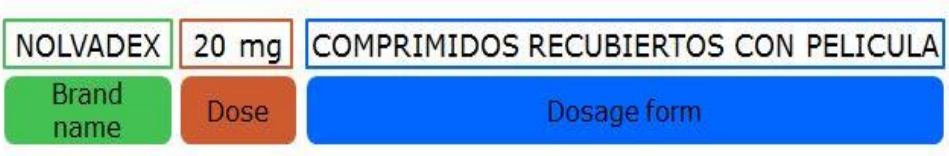

Figure 19: Summary of brand name creation

After the review of the CIMA resource and the creation of the drug part of the dictionary, the final numbers obtained are the following:

RESOURCE	TOTAL
Generic drugs from CIMA	2,228
Brand drugs from CIMA	3,662
Total drugs from CIMA:	5,890

Table 12: Number of drugs in the dictionary

4.2.3. Adverse Event dictionary

The two selected resources regarding the adverse events were MedDRA and Vademecum. Due to the structure of the first one, it was chosen to be the leading one in the dictionary.

The two lower levels from MedDRA (PT - Preferred Terms and LLT - Lowest Level Terms) were extracted from the resource. Level 5 (last level, PPT) is composed of specific adverse events (i.e. *Dolor de cabeza - Headache*), whereas in level 4 (PT) more general adverse events are found (i.e. *Cefalea - Cephalgia*). 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.

There are a total of 72.072 Lowest Level Terms, and 20.307 Preferred Terms. These last ones are included in the first ones (thus, every PT has at least one alias - itself). Therefore, with the implementation decision, the dictionary has 20307 entries (PT) for adverse events, and a total of 72072 aliases (LLT).

For example, *dolor hepático* (Liver pain), with MedDRA code 10019705, is a Preferred Term from the level 4 in MedDRA. In the dictionary it will be a new entry for an adverse event, with ID equal to the code in MedDRA. In the level 5, as Lower Level Terms associated to *Dolor hepático*, there is a list of terms, such as itself, *dolor hepático* (code 10019705), *dolor de hígado* (code 10024703) and *dolor hepatobiliar* (code 10057960). These terms will be the aliases.

As a matter of fact, after the creation of the adverse event part of the dictionary, the final numbers in our resource were:

RESOURCE	TOTAL
Adverse events from MedDRA	72,072
of which as entries	20,307
of which as aliases	72,072
Total adverse events:	72,072

Table 13: Number of adverse events in the dictionary

4.3. Construction of gazetteers

By analyzing the information from these resources, we found that none of them contained all of the drugs and adverse events. Patients usually use lay terms to describe their symptoms and their treatments. Unfortunately, many of these lay terms are not included in the above mentioned resources. Therefore, we decided to integrate additional information from other resources devoted to patients to build a more complete and comprehensive dictionary. There are several online websites that provide information to patients on drugs and their side effects in Spanish language. For example, Vademecum contains information about drugs and their side effects. This website allows users to browse by generic or drug name, providing an information leaflet for each drug in a HTML page. Since these leaflets are unstructured, the extraction of drugs and their adverse effects is a challenging task. While drug names are often located in specific fields (such as title), their adverse events are usually descriptions of harmful reactions in natural language. We developed a web crawler to browse and download pages related to drugs from Vademecum since this website provided an easier access to its drug pages than others, such as MedLinePlus.

With the Textalytics dictionary complete, the main resources were integrated and could be annotated by the system by means of the Textalytics Topic Extraction API.

Nevertheless, there were other resources to integrate, plus extra information which could be important in order to enlarge the scope for annotation. In order to achieve this goal, different gazetteers were implemented.

A gazetteer can be defined as a set of lists which contains entities (in this case, drugs, generic ingredients, families of drugs and adverse events), which will be annotated by GATE if there are occurrences in a text.

Drug families - WHO ATC system gazetteer

Users may refer to a specific drug by its brand name (*Espidifén)*, its active ingredient (*Ibuprofeno)* or in a more general way, by its family name (*antiinflamatorios)*. The first two options are included in the drug part of the dictionary, whereas the third possibility was not part of the system.

The ATC code (Anatomical Therapeutic Chemical Classification System) is a pharmaceutical coding system which is organized by therapeutic groups, and is controlled by the WHO Collaborating Center for Drug Statistics Methodology (WHOCC).

The system is structured in five levels, being the last one the generic ingredients, and the fourth one the chemical/therapeutic/pharmacological subgroups. By obtaining the subgroups from all the 14 anatomic groups (first level), and organizing them in families of drugs, a gazetteer with 466 families was created.

Drugs and Adverse events – Vademecum gazetteer

The main limitation of CIMA is that it only provides information about drugs authorized in Spain. That is, CIMA does not contain information about drugs approved only in Latin America.

Nevertheless, a total of 4238 drugs were obtained from Vademecum, and then compared with the ones in CIMA in order to filter the ones which were included in both resources, and to create a gazetteer with the ones from Vademecum which were not part of CIMA. A gazetteer with 1237 drugs was compiled with these drugs.

On the other hand, the main limitation of MedDRA is that that its contents are not in lay language, and therefore there are not always found in a user's post.

Thus, the 2793 adverse events which were gathered with the web crawler from Vademecum, which were more patient-oriented, were also included in the gazetteer.

After the extraction and implementation of all the drugs and adverse events which appeared in the resources that we studied, the final numbers in terms of entities were:

RESOURCE	TOTAL
Generic drugs from CIMA	2,228
Brand drugs from CIMA	3,662
Drug group names from the ATC system	466
Drug names (which are not in CIMA) from Vademecum	1,237
Total Drugs:	7,593

Table 14: Number of drugs in total

RESOURCE	TOTAL
Adverse events from MedDRA	72,072
Adverse events from Vademecum (which are not in MedDRA)	2,793
Total adverse events:	74,865

Table 15: Number of adverse events in total

5. Evaluation

5.1. Metrics for information extraction

In order to be able to evaluate the obtained results it is critical to choose clear, reproducible and easily understood evaluation metrics. The first thing before introducing the metrics is to define a structure known as confusion matrix or contingency table. This is divided into four categories:

- *True Positives* (*TP*): Examples correctly labeled as positives
- *False Positives* (*FN*): Negative examples incorrectly labeled as positives
- *True Negatives* (*TN*): Negative examples correctly labeled as positives
- *False Negatives* (*FN*): Positive examples incorrectly labeled as negatives

	Predicted Positive	**Predicted Negative**
Actual Positive	*TP*	*FN*
Actual Negative	*FP*	*TN*

Table 16: Confusion matrix

Table 16 shows the structure of the confusion matrix, which is a table in which each column represents the number of predictions of each class, whereas the rows represent the instances of the real class. Therefore, each element shows the number of examples for which the actual class is the row and the predicted class is the column.

True positives and false negatives are mutually related, as it happens with false positives and true negatives. We will define N^+ as the total number of positive examples, whereas N^- will be the total number of negative examples. Therefore, it is understandable that *TP* and *FN* are complementary labels for N^+ and the same way *TN* and *FP* are for N^-. It is important to mention that *TP* and *FP* are the only two of all the numbers *TP*, *FP*, *TN*, and *FN* which are independent.

$$N^+ = TP + FN \tag{5.1}$$

$$N^- = TN + FP \tag{5.2}$$

In the following list we will define several metrics based on the confusion matrix:

- *True Positive Rate* (*TPR*): Measures the fraction of positive examples which is correctly labeled. *TPR* is also referred to as *sensitivity* or *recall*. This score is typically used to evaluate the performance of medical tests.

$$TPR = Sensitivity = Recall = \frac{TP}{TP+FN} = \frac{TP}{N^+} \tag{5.3}$$

- *False Positive Rate* (*FPR*): Measures the fraction of negative examples that is wrongly classified as positive.

$$FPR = \frac{FP}{TN+FP} = \frac{FP}{N^-} \tag{5.4}$$

- *True Negative Rate* (*TNR*): Measures the fraction of negative examples that is correctly classified as negative. *TNR* is also referred to as *specificity*. In medical applications, *FPR* is replaces with *TNR*.

$$TNR = Specificity = \frac{TN}{TN+FP} = \frac{TN}{N^-} \tag{5.5}$$

- *False Negative Rate* (*FNR*): Measures the fraction of positive examples that is wrongly classified as negative.

$$FNR = \frac{FN}{FN+TP} = \frac{FN}{N^+} \tag{5.6}$$

From these equations we make it clear that:

$$TPR + FNR = 1 \tag{5.7}$$

$$FPR + TNR = 1 \quad (5.8)$$

As we are dealing with an information extraction system, whose performance is normally evaluated with the recall (equation 5.3) and precision values, we must introduce the definition of the last one:

- *Precision*: The proportion of detected examples that are actually positive examples

$$Precision = \frac{TP}{TP+FP} \quad (5.9)$$

An ideal information extracting system is the one where *FN* = 0 and *FP* = 0. Precision and recall measures stand in opposition to one another. When precision increases, recall usually decreases, and vice versa.

- *F-measure*: Defined by the harmonic (weighted) average between precision and recall, where the parameter *beta* indicates a relative weight of precision with respect to the recall. When β = 1, the balanced *F-score* (F_1) is such that recall and precision are evenly weighted. Meanwhile, when β = 2 (F_2), an overall performance is achieved, in this case, giving more importance to the precision (double the recall).

$$F(\beta) = \frac{(1+\beta^2)PR}{\beta^2 P+R} \quad (5.10)$$

5.2. Results

We evaluated the system in two steps on the corpus annotated with drugs and adverse events. The first time we measured the results it was when only the dictionary was created, but no extra information was induced by the gazetteers. The second evaluation took place with the complete system as it is known throughout this document.

First evaluation. Only dictionary with drugs and adverse events

	PRECISION	RECALL	F-MEASURE
Drugs	0.69	0.47	0.56
Adverse Events	0.83	0.37	0.51

Table 17: First evaluation results

The results of this first evaluation show a precision of 69% for drugs and 83% for adverse events, and a recall of 47% for drugs and 37% for adverse events. This concluded in an F-score of 56% for drugs and 51% for adverse events.

The results regarding drugs were lower than expected, and that is what made us think about possible improvements, which came in form of gazetteers. On the one hand, we noticed that the families of drugs were a very common issue that it was easily solved with the ATC code information. Moreover, the integration of a resource like Vademecum gave us over 1.000 extra drugs. On the other hand, the low results provided by the adverse events were more expected. The lay language used in social media will make this result very difficult to improve. Anyway, with the integration of Vademecum in the gazetteer, more adverse events were included in the system's resources.

Second evaluation. Dictionary and gazetteers (complete system)

	PRECISION	RECALL	F-MEASURE
Drugs	0.88	0.80	0.84
Adverse Events	0.85	0.56	0.67

Table 18: Second evaluation results

The second and final evaluation, after integrating the gazetteers into the system showed a high improvement. In this case we can see a precision of 88% for drugs and 85% for adverse events, and a recall of 80% for drugs and 56% for adverse events. This concluded in an F-score of 84% for drugs and 67% for adverse events.

We could see, as we were expecting, that the integration of the family of drugs gazetteer improved considerably the results for the drugs. Furthermore, the improvement in the adverse events part of the dictionary in addition to the extra events obtained from Vademecum, which were written in a more patient-oriented way, was clearly visible in the increase of the adverse events results.

Evaluation comparison

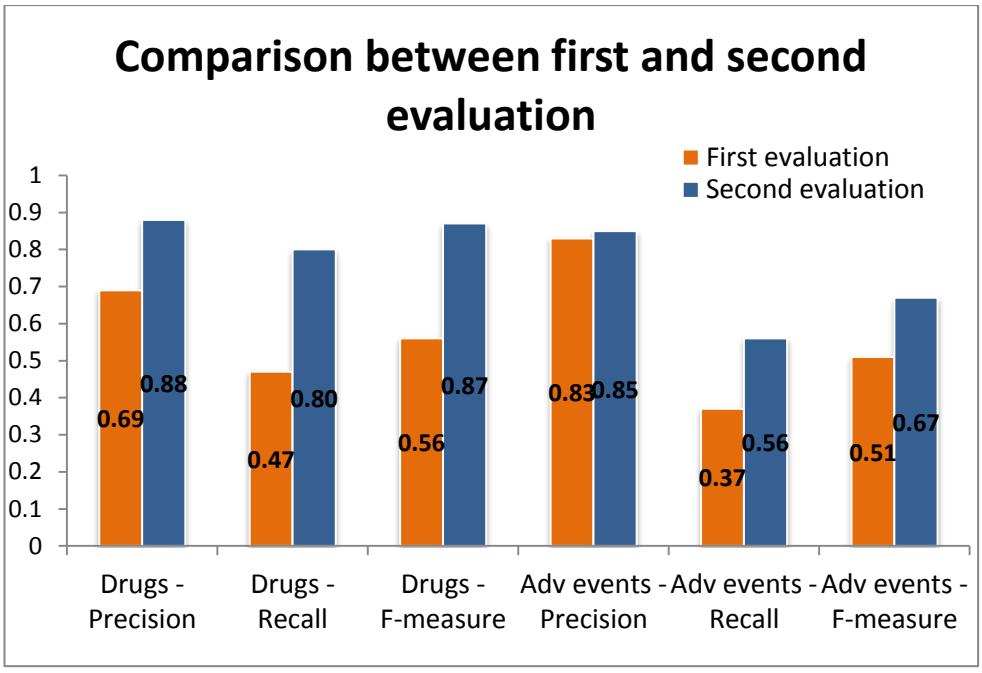

Figure 20: Comparison between first and second evaluation

We can see a very important increase in all the results regarding drugs. As a matter of fact, in the F-score there is an uplift of 31 p.p., what means an improvement of over 55% with respect to the first evaluation. What is more, the biggest upgrade is found in the drug's recall, with an uplift of 33 p.p.

The improvement in terms of adverse reactions is not as high due to the fact that the results for precision in the first evaluation were already very high. Anyway, the uplift of 19 p.p. in terms of recall increases the F-score in 16 p.p. up to a satisfactory 67%.

5.3. Error analysis

We performed an analysis to determine the main sources of error in the system. A sample of 50 user comments were randomly selected and analyzed.

Adverse events error analysis

False Negatives

Regarding the detection of adverse events, the main source of false negatives was the use of colloquial and lay expressions to describe an adverse

event. Expression like "*me deja ko" (it makes me KO)*, "*mi vida no va a ningún lado" (my life is going nowhere)*, or "*me cuesta más levantarme" (it's harder for me to wake up)* were used by patients in order to express their adverse events. These phrases are not included in our resources. The idiomatic expressions along with the desire of emphasizing statements make it very difficult to create a suitable lexicon.

The second highest cause of false negatives for adverse events was due to the different lexical variations of the same adverse event. For instance, "*depresión" (depression)* is a term included in our dictionary, but its lexical variations, like for example "*deprimido" (depressed - masculine), "deprimida" (depressed - feminine), "me deprimo" (I get depressed), "depresivo" (depressive), "deprimente" (depressing)* are not, and therefore are not detected by our system.

The third largest cause is spelling mistakes. We can see an example with "*hemorrajia"*, which is an incorrect way of writing "*hemorragia"* (*hemorrhage)*. Many users have great difficulty in spelling unusual and complex technical terms like the ones appearing in medicine.

The last important error source was the use of abbreviations ("*depre"* is an abbreviation for "*depression"*), which also produces false negatives.

False Positives

False positives for adverse events were mainly due to the inclusion of MedDRA terms referring to procedures (such as therapeutic, preventive or laboratory procedures) and tests in our dictionary. MedDRA includes terms for diseases, signs, abnormalities, procedures and tests. We should have not included those terms referring to procedures and tests since they do not represent adverse events. In this case we are talking of words such as "*prevención" (prevention), "análisis" (analysis)* or *"cirugía" (surgery)*.

Summary

FALSE NEGATIVES (FN)	
Cause	**Example**
Lay expressions	*Me deja KO* (it makes me KO) *Mi vida no va a ningún lado (*my life is going nowhere*)* *Me cuesta más levantarme (*it's harder for me to wake up*)*
Lexical variability	*Depresión* (depression) *Deprimido* (depressed - masculine) *Deprimida* (depressed - feminine) *Me deprimo* (I get depressed) *Depresivo* (depressive) *Deprimente* (depressing)
Spelling mistakes	Hemorrajia [*hemorragia*] (hemorrhage)
Abbreviations	Depre [*depresión*] (depression)

Table 19: False negatives in the AdverseEvent recognition task

FALSE POSITIVES (FP)	
Cause	**Example**
Non-adverse events terminology	*Prevención* (prevention) *Análisis* (analysis) *Cirugía* (cirugía)

Table 20: False positives in the AdverseEvent recognition task

Drugs error analysis

False Negatives

The main source of false negatives for drugs seems to be that users often misspelled drug names. Some generic and brand drugs have complex names for patients. Some examples of misspelled drugs are *Avilify* (*Abilify*) or *Rivotril (Ribotril).*

Another important cause of false negatives was due to the fact that our dictionary does not include drugs approved in other countries than Spain (for example, *Clorimipramina*, *Ureadin* or *Paxil*). However, ForumClinic has a large

number of users in Latin America. It is possible that these users have posted comments about some drugs that have only been approved in their countries.

The third largest source of errors was the abbreviations for drug families. For instance, *benzodiacepinas (benzodiazepine)* 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. This combination of drugs is not registered in the resources (CIMA and Vademecum) used to create our dictionary.

False Positives

Most false positives for drugs were due to a lack of ambiguity resolution. Some drug names are common Spanish words such as "*Allí"* (a slimming drug) or "*Puntual*" (a laxative). These terms are ambiguous and resolve to multiple senses, depending on the context in which they are used.

Similarly, some drug names such as "alcohol" or "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*).

Summary

FALSE NEGATIVES (FN)	
Cause	**Example**
Drug name misspelling	*Avilify [Abilify]* *Rivotril [Ribotril]*
Drugs outside Spain	*Clorimipramina* *Ureadin* *Paxil*
Drug famililes abbreviations	Benzos (*Benzodiacepinas)* [Benzodiazepine]
Acronyms for drug combinations	*FEC*: *Fluorouracil, Epirubicin* and *Cyclophosphamide*

Table 21: False negatives in the DrugName recognition task

FALSE POSITIVES (FP)	
Cause	**Example**
No ambiguity resolution	*Allí* (There) [Sliming drug] *Puntual* (Punctual) [Laxative]
Drug family as adjective	*Sedante* (Sedative) *Antidepresivo* (Antidepressant) [*La pastilla que me mandó me ha causado un efecto sedante*]

Table 22: False positives in the DrugName recognition task

6. Budget

6.1. Project description

Author: Ricardo Revert Arenaz

Department: Informatics

Title: Detecting drugs and adverse events from Spanish health social media streams

Duration: Started September 9th 2013 and ended January 24th 2014. Totally, 560 hours were needed for the realization of this project.

6.2. Costs calculation

Wages of personnel

Project carried out by a single person, which took responsibility of all the tasks. In the next table the breakdown of personnel is shown:

STAGE	COST/HOUR	TOTAL HOURS	TOTAL COST
Analysis	€ 90,00	75	€ 6.750,00
Design		60	€ 5.400,00
Coding		150	€ 13.500,00
Tests		110	€ 9.900,00
Experiments		75	€ 6.750,00
Documentation		90	€ 8.100,00
TOTAL		560	€ 50.400,00

Table 23: Wages of personnel

Equipment costs

For the realization of this project, a laptop was used during the entire duration, as well as a desktop computer was used for tests and experiments for one month.

CONCEPT	UNIT	TOTAL COST
Laptop	1	€ 785,00
Testing desktop computer	1	€ 580,00
TOTAL		€ 1.365,00

Table 24: Equipment costs

Software costs

The cost of the operating system of both computers is included in the table above as it was included with the hardware. Furthermore, GATE is a free open source software so it has no cost, and Textalytics Professional API has a cost of 149 €/month (with I.V.A). Regarding the documentation stage, Microsoft Office 2010 was included with the operating system.so it is included in the table above.

CONCEPT	UNIT	TOTAL COST
Operating System	2	€ 0,00
GATE	1	€ 0,00
Textalytics Professional API	1	€ 123,00
TOTAL		€ 123,00

Table 25: Software costs

Expendable and other costs

The cost of material such as paper, printer ink and other expenses which have not been considered in the aforementioned concepts are:

CONCEPT	TOTAL COST
Paper	€ 30,00
Printer ink	€ 20,00
Other expenses	€ 90,00
TOTAL	€ 140,00

Table 26: Expendable and other costs

Indirect costs

In order to do an approximation of the indirect costs, we are going to suppose an indirect cost rate of 20%.

6.3. Budget

Next, the breakdown of the total budget of the project based on all the costs above explained:

CONCEPT	TOTAL COST
Wages of personnel	€ 50.400,00
Equipment costs	€ 1.365,00
Software costs	€ 123,00
Expendable and other costs	€ 140,00
Indirect costs	€ 10.405,60
Value-add tax (I.V.A - Impuesto sobre el Valor Añadido – Tipo general del 21%)	€ 13.111,06
TOTAL	**€ 75.544,66**

Table 27: Total budget

Therefore, the total cost of the budget is of **SEVENTY FIVE THOUSAND FIVE HUNDRED FOURTY FOUR EUROS AND SIXTY SIX CENTS** (€ 75.544,66), including the value-add tax corresponding to the general type of 21%

7. Conclusions and future work

7.1. Accomplished objectives

In this research, we created the first Spanish corpus of health user comments annotated with drugs and adverse events. This corpus is now available for research. In this work, we focused only on detecting the mentions of drugs and adverse events, and not on the relationships among them.

Furthermore, we also implemented the first system working in Spanish for the detection of drugs and adverse events. Nevertheless, to the best of our knowledge there are now other systems working in Spanish detecting drugs and adverse events, but not oriented to health social media streams.

Moreover, there is pipeline for entities detection now available online[37], where the system can be tested.

The creation of a gold-standard corpus with comments from health related social media is of a high importance regarding the future development and improvements in information extraction techniques in Spanish.

[37] http://163.117.129.57:8090/gate/

In terms of performance, the developed system turned out to have very successful results. The F-score for the drugs was of 87% and for the adverse events of 67%, leaving an unweight average of 77%. The fact that the results for drugs are higher than the ones for adverse events has a straight forward explanation. Drugs are a finite source, where medicines have a name and a reduced way of referring to them. This makes it possible to create a more precise dictionary regarding drug names and families, and leaving as a main source of error misspellings and abbreviations. On the other hand, the ways to refer to adverse events are uncountable. There are as many ways as patients expressing them on social media. The lay language is the most common when dealing with forums, and it is almost impossible to obtain a dictionary which includes all of them. Therefore, we will later take a look at other ways of improving the system.

Another important issue to point out is the outstanding improvement committed over the system after the first evaluation. With still quite good results in precision (69% for drugs and 83% for adverse events), the poor performance in recall (47% for drugs and 37% for adverse events) made us reconsider the direction the system was taking and found the answer to our problems. This was the creation of the gazetteers, which would give us more lay language, and a reduced cleanup of the dictionary, which made the system obtain much better results. In fact, as it was said in Chapter 5, the improvement in the drug's F-score was of 31 p.p. (56% vs. 87%), whereas in the adverse event's one was of 16 p.p. (51% vs. 67%), concluding in an unweight average improvement of 23.5 p.p. (53.5% vs. 77%).

Despite the good results, the system can still reach a better performance. The main sources of errors which can be found when evaluating a health social media based corpus are the lay expressions and spelling and grammatical mistakes, as well as the use of abbreviations and the lexical variations that words offer. With the purpose of moving on and improving the system, we will comment the future guidelines to follow.

It is interesting to comment that with the content of this study, a research article (Segura-Bedmar et al., 2014 [28]) was introduces in the 5th International

Workshop on Health Text Mining and Information Analysis (Louhi[38]), which took place in the 14th Conference of the European chapter of the Association for Computational Linguistics (EACL[39]), in Gothenburg, Sweden, were it was selected and published in the proceedings[40] for Louhi 2014.

7.2. Future work

To begin, the main objective would be to improve on the spots where the system is obtaining worse performance. Therefore, taking a look at the main sources of error, we could obtain the first future guidelines.

The first and most obvious step would be integrating more resources in the system. To date, we have CIMA for drugs and MedDRA for adverse events, with support of gazetteers compiled from resources such as Vademecum or the WHO ATC code. The more integrated resources, the better the performance is expected to be.

As for the main source of error in this kind of text, the lay language and colloquial expressions, the main improvement can come from the creation of a lexicon formed by this kind of vocabulary and expressions. The larger the lexicon with lay expressions, the higher the improvement in the results will be.

Another improvement which could show very good results would be the introduction of nominalization for lexical variations. In this way, and following the example in Chapter 5, by including the word *depresión* in the dictionary, the system would detect as many lexical variations as a patient can express. Thus, terms such as *deprimido* (depressed), *me deprimo* (I get depressed), *depresivo* (depressive) or *deprimente (*depressing*)* would no longer imply false negatives.

Along with the lexical variations and the lay expressions, we have a very important source of error like the spelling mistakes. In order to suppress this issue, the introduction of advanced matching methods would suppose a significant improvement in the results. Computing distances between two words

[38] http://dsv.su.se/en/research/research-areas/language/description/louhi-2014-1.154970
[39] http://eacl2014.org/
[40] http://www.aclweb.org/anthology/W/W14/#1100

(the correct one and the one with the spelling mistake), this problem can come to an end.

Moreover, in future work, we plan to extend the system to detect the relations between drugs and their side effects. Furthermore, we would like to identify their indications and beneficial effects.

Other possible guidelines to future work could be developing a multilingual system which is capable of being applied to different languages. The positive starting point would be the fact that MedDRA's contents exist in different languages, which makes it possible to relate expressions by a code that they share. Of course, other resources from different languages should be integrated, both for drugs and adverse events.

Finally, a differentiated guideline for the future would be the development of a corpus and a system for other kind of documents in Spanish rather than social media, as for example, drug packet inserts. Despite seeming unrelated to this project, the majority of the resources could be shared, and the discovery of different ones could contribute in an improvement for both systems.

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases,* 1215:487-499.

[2] Alan R Aronson and Francois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229-236.

[3] Alexandra Balahur. 2013. Sentiment Analysis in Social Media Texts. *WASSA 2013*, 120.

[4] David W. Bates, R Scott Evans, Harvey Murff, Peter D. Stetson, Lisa Pizziferri and George Hripcsak. 2003. Detecting adverse events using information technology. *Journal of the American Medical Informatics Association*, 10(2):115-128.

[5] Adrian Benton, Lye Ungar, Shawndra Hill, Sean Hennessy, Jun Mao, Annie Chung, Charles E. Leonarda and John H. Holmes. 2011. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics*, 44(6): 989-996.

[6] Jiang Bian, Umit Topaloglu and Fan Yu. 2012. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, 25-32.

[7] CA. Bond and Cynthia L. Raehl. 2006. Adverse drug reactions in United States hospitals. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 26(5):601-608.

[8] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychol Meas*, 20:37e46.

[9] Ronald A. Fisher. 1922. On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87-94.

[10] Flavien Bouillot, Phan N. Hai, Nicolas Béchet, Sandra Bringay, Dino Ienco, Stan Matwin, Pascal Poncelet, Mathiue Roche and Maguelonne Teisseire. 2013. How to Extract Relevant Knowledge from Tweets? *Communications in Computer and Information Science*.

[11] Carol Friedman. 2009. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In *Artificial Intelligence in Medicine.* LNAI 5651:1-5.

[12] Graciela H. Gonzalez, Matthew L Scotch and Garrick L Wallstrom. Mining Social Network Postings for Mentions of Potential Adverse Drug Reactions. HHS-NIH-NLM (9/10/2012 - 8/31/2016).

[13] Harsha Gurulingappa, Abdul Mateen-Rajput and Luca Toldo. 2012. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics,* 3(1):15.

[14] Harsha Gurulingappa, Luca Toldo, Abdul Mateen-Rajput, Jan A. Kors, Adel Taweel and Yorki Tayrouz. 2013. Automatic detection of adverse events to predict drug label changes using text and data mining techniques. *Pharmacoepidemiology and drug safety*, *22*(11):1189-1194.

[15] A Herxheimer, MR Crombag and TL Alves. 2010. Direct patient reporting of adverse drug reactions. A twelve-country survey & literature review. *Health Action International (HAI). Europe.* Paper Series Reference 01-2010/01.

[16] Shawndra Hill, Raina Merchant and Lile Ungar. (2013). Lessons Learned About Public Health from Online Crowd Surveillance. *Big Data*, 1(3):160-167.

[17] George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc*, 12:296e8.

[18] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal and Alfonso Valencia. 2013. Overview of the chemical compound and drug name recognition (CHEMDNER) task. In *BioCreative Challenge Evaluation Workshop,* 2:2-33.

[19] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars J. Jensen and Peer Bork. 2010. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(343):1-6.

[20] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing.* 117-125. Association for Computational Linguistics.

[21] Anne J. Leendertse, Antoine C. Egberts, Lennar J. Stoker, & Patricia M.L.A. van den Bemt. 2008. Frequency of and risk factors for preventable medication-related hospital admissions in the Netherlands. *Archives of internal medicine*, 168(17), 1890.

[22] Qi Li, Louise Deleger, Todd Lingren, Haijun Zhai, Megan Kaiser, Laura Stoutenborough Anil G Jegga, Kevin B Cohen and Imre Solti. 2013. Mining FDA drug labels for medical conditions. *BMC medical informatics and decision making*, 13(1):53.

[23] Mark McClellan. 2007. Drug Safety Reform at the FDA-Pendulum Swing or Systematic Improvement? *New England Journal of Medicine*, 356(17):1700-1702.

[24] Silvio Moreira, Joao Filgueiras, Bruno Martins, Francisco Couto and Mario J. Silva. 2013. REACTION: A naive machine learning approach for sentiment classification. In *2nd Joint Conference* on. *Lexical and Computational Semantics, 2:490-494.*

[25] Melanie Neunerdt, Michael Reyer and Rudolf Mathar. 2013. A POS Tagger for Social Media Texts trained on Web Comments. *Polibits*, 48:59-66.

[26] Azadeh Nikfarjam and Graciela H. Gonzalez. 2011. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings,* 2011:1019-1026. American Medical Informatics Association.

[27] Isabel Segura-Bedmar, Paloma Martınez and María Herrero-Zazo. 2013. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). 3206(65): 341-351.

[28] Isabel Segura-Bedmar, Ricardo Revert and Paloma Martínez. 2014. Detecting drugs and adverse events from Spanish social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi),* 106-115

[29] Cornelis S. van Der Hooft, Miriam CJM Sturkenboom, Kees van Grootheest, Herre J. Kingma and Bruno HCh Stricker. 2006. Adverse drug reaction-related hospitalisations. *Drug Safety*, 29(2):161-168.

[30] Hamish Cunningham. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223-254.

[31] M Rawlins. 1995. Pharmacovigilance: paradise lost, regained or postponed? The William Withering Lecture 1994. *Journal of the Royal College of Physicians of London*, 29(1): 41-49.

[32] Sunghwan Sohn, Jean-Pierre A. Kocher, Christopher G. Chute and Guergana K. Savova. 2011. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, 18(Suppl 1): i144-i149.

[33] Özlem Uzuner, Imre Solti and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*. 17(5):514-518.

[34] Rong Xu and QuanQiu Wang. 2013. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC Bioinformatics*, 14(1):181.

[35] Karin Wester, Anna K. Jönsson, Olav Spigset, Henrik Druid and Staffan Hägg. 2008. Incidence of fatal adverse drug reactions: a population based study. *British journal of clinical pharmacology*, 65(4):573-579.

[36] Yonghui Wu, Joshua C. Denny, S. Trent Rosenbloom, Randolph A. Miller, Dario A. Giuse, and Hua Xu. 2012. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 997–1003.