PhUSE 2017

Paper DH01

Patient-generated Health Data (Social Media) is a Potential Source for ADR Reporting

Erwan Le Covec, Keyrus Biopharma, Lasne, Belgium
Essam Ghanem, Keyrus Biopharma, Lasne, Belgium
Stéphane Chollet, Keyrus Biopharma, Levallois-Perret, France

ABSTRACT

Social media in Pharmacovigilance science has a crucial influence and visible role on public health protection, evaluating the medicines products' benefits versus potential risk.

Social media are computer-mediated technologies that facilitate the creation and sharing of information and an important data source for adverse drug reaction (ADR) and can help to reduce under-reporting in product post-marketing phase. Their application is subjected to some challenges as the data can be inconsistent, unstructured and represent a huge volume.

This paper proposes an automated approach that enables effective ADR detection and extraction from social media, focusing on Twitter® and Reddit® forum. Automated solution to manage unstructured received data in a huge volume is proposed (relevant unpublished data will be provided). This method, which is still under evaluation, will effectively extract ADR in targeted social media sources and is compliant for multiple ones. Facing social media data 'inconsistency', it requires implementation of manual revision.

INTRODUCTION

Social media is a form of electronic communication through which users create online communities to share information, ideas, personal messages, and other content[1]. Due to the inevitable impact of the social media, people tend to have a stronger presence on the Web and do not hesitate to share information which would have been kept private several years ago (2006, the year Twitter was created[2] and Facebook® started to have a significant number of users[3]). The importance of digital media sources in Pharmacovigilance such as Twitter, Facebook, health-related forums and search queries to extract Adverse Events (AE) information was recently highlighted in the literature[7,8,9].

It is now possible to find Adverse Drug Reaction (ADR) associated with a treatment mentioned by patients seeking advice on the received medicines whether Prescription Only Medicine (POM) or Over the Counter (OTC). Patient-generated health data in social media are considered as a potential source for ADR reporting that are currently receiving significant attention in the Pharmacovigilance field. Different approaches for ADR data extraction from social media were proposed in literature facing the problem of data inconsistency[4, 5]. The Innovative Medicines Initiative (IMI), is also putting some efforts on the subject by working on technical tools for data mining on social media websites. But such unstructured data sources render the safety data triage somewhat difficult with a remarkable complexity (ex: vocabulary inconsistency). In addition, handling safety data from multiple sources in huge amounts needs justification and is not optimized for databases using a relational schema.

The present paper is proposing an automated method to effectively extract, detect and store ADRs from Twitter and the website Reddit.

CONTEXT

PHARMACOVIGILANCE NEEDS

The Pharmacovigilance science is related to the collection, detection, assessment, monitoring, and prevention of adverse events (AEs) with pharmaceutical products. Safety signal is information on a new or known AE (whose frequency increases) that may be caused by a medicine and requires further investigation, a wide range of sources can be involved to detect them. According to the Guidelines on Good Pharmacovigilance Practices (GVP), social media can be a source of potential valid Individual Case Safety Reports (ICSRs) because they allow patients and healthcare professionals to communicate ADRs and thus can be used as sources for new signals identification[6].

Detecting AEs from social media is not part of the Pharmacovigilance process as they are not classified as signals yet, but could help for future signal detection.

EXISTING RESEARCHES AND OUTCOMES QUESTION

Several researches have already been made in the past about ADR detection on social media. Exploring Twitter as a resource for the automatic extraction of ADRs was the aim of "*Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions*"[7]. The authors tested the effectiveness of the commonly used lexicon-based methods which were able to detect more than 50% of the ADRs. However, several limitations were found: some tweets needed manual detection and some drugs were not well represented on Twitter.

Using a health-related social network was tried in "*Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments*"[8]. The authors presented a system to automatically extract mentions of ADRs from user reviews in the health-related social network DailyStrength®. The system applied association rule mining on a set of annotated comments to extract expressions about adverse effects. The extraction results were slightly better than with the lexicon-based methods but very dependent on the number of annotated data which is time consuming.

To address some limitations due to a single source use, the authors of "*Pharmacovigilance from Social Media*"[9] assessed data from both Twitter and DailyStrength. With a new approach based on machine learning and clustering, the authors managed to achieve a better ADR detection compared to the lexicon-based methods but didn't try any normalization step.

Existing research highlights that using only one source of data leads to limitation. Some data are not well represented in some sources, thus creating statistics bias. None of the aforementioned researches has considered that data normalization via medical codification might have an impact on the data obtained from various sources. In view of the aforementioned literature data, the present paper is proposing a methodology for structured data collection, enabling automated ADR detection and extraction from two unstructured data collection system.

PROPOSED METHOD

The proposed method aims to detect, normalize and store ADRs in text extracted from two social media: Twitter and Reddit, the method being compliant as well for other kinds of Internet sources. The use of more than one data source, should increase the representation of some drugs as well as the global number of data and allow a more effective ADR detection compared with the use of one data source only.

Detection is done by the lexicon-based methods which consist in searching if a word of a lexicon is present in a text. A normalization step is performed to code the drugs and the ADR detected with, respectively, the WHO Drug Dictionary Enhanced (WHO-DDE) and the Medical Dictionary for Regulatory Activities (MedDRA).
After each main step, the data will be stored in a non-relational database due to the particular structure of the data.

DATA SOURCES

SOCIAL MEDIA PRESENTATION

Social media are technologies that facilitate the creation and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks[10]. There are various kinds of social media and their common features are:

- Interactivity with Internet-based applications.
- User-generated content, such as text posts, comments, photos or videos. These data generated through online interaction are the core of social media.
- User profile designed and maintained by the social media organization.
- Online social networking by connecting users' profiles with other individuals or groups.

User-generated contents are often personal information users want to share with their contacts or with other social media users. They can carry some valuable data if analyzed correctly resulting in sentiment analysis, influence measurement, trends or pattern detection[11]. According to some studies, 30% of adults are likely to share information about their health on social media sites with other patients[12].

The sources were selected because:

- Twitter is a famous social network, used for many purposes, from news to personal data sharing. It is used in many countries so it can provide a lot of personal information for many subjects like medical related topics.
- Reddit is a famous social news aggregation, web content rating, and discussion website. It behaves like a forum and focuses on connecting users who share similar interest and have special categories for drugs usage[13].

TWITTER

Twitter is an online news and social networking service where users post and interact with messages restricted to 140 characters. Messages are called tweets. In April 2017, it was the 10th most used social network with 319,000,000 active users[14] and an average of 58,000,000 messages per day[15]. This potential gold mine of information for researchers interested in studying population trends can be easily used because a part of the Twitter data is publicly available and easily accessible[16].

Twitter data are provided in JavaScript Object Notation (JSON). It is an open-standard file format which represents information with tags allowing the file to be human-readable (see Figure 1).

```
{
  "firstName": "Erwan",
  "lastName": "Le Covec",
  "isAlive": True,
  "age": 25
}
```

```
{
  "date": datetime.datetime(2017, 6, 19,
15, 57, 22),
  "retweeted": False,
  "text": u"I took Gabapentin for my bach
pain and now I have nausea…",
  "user_id": 704703338261,
  "user_location": u"USA & Switzerland",
  "user_name": u"erwanlc",
  "drug": u"Gabapentin"
}
```

Figure 1: Examples of JSON object.

This data format is very common for browser/server communication and is language-independent.

WEBSITE REDDIT

Reddit is one of the most visited discussion website in the world (#9). In 2017, it had 542 million monthly visitors. On this social network, users can connect with others with the same interest by visiting the categories they want from video games to politic through medicine and drugs usage[13].

There are some special sections to talk about treatments, drugs and eventual side effects which make the website very useful for ADRs detection. In these categories user can share pictures but also share their experiences, the treatments they received or currently in use as well as their current state of mind and physical shape.

To get the targeted information, Reddit proposes also an API but using it only allow finding text with a specific keyword in it. As some discussions are about specific drugs but don't contain its mention in every post, some useful information can be lost and thus Hypertext Markup Language (HTML) of the website needs to be analyzed. HTML is the standard markup language for web pages creation which provides a framework but due to the amount of information, valuable information can be difficult to retrieve. It is a challenge as most of the websites have different HTML structure.

```
<div class=" thing id-t1_dn8ymcn noncollapsed comment " id="thing_t1_dn8ymcn" onclick="click_thing(this)" data-fullname="t1_dn8y" data-type="comment" data-subreddit="benzodiazepines" data-subreddit-fullname="t5_2s4" data-author="erwanlc" data-author-fullname="t2_9dq8"><p class="parent"><a name="dn8ymcn"></a></p><div class="midcol unvoted"><div class="arrow up login-required access-required" data-event-action="upvote" role="button" aria-label="vote positif" tabindex="0"></div><div class="arrow down login-required access-required" data-event-action="downvote" role="button" aria-label="vote négatif" tabindex="0"></div></div><div class="entry unvoted"><p class="tagline">
```

```
<a href="javascript:void(0)" class="expand" onclick="return togglecomment(this)">[–]</a><a href="https://www.reddit.com/user/erwanlc" class="author may-blank id-t2_9dq80du">erwanlc</a>
<span class="userattrs"></span> <span class="score dislikes" title="2">2 points</span><span class="score unvoted" title="3">3 points</span><span class="score likes" title="4">4 points</span> <time title="Wed Sep 20 04:44:18 2017 UTC" datetime="2017-09-20T04:44:18+00:00" class="live-timestamp">il y a 10 heures</time> <a href="javascript:void(0)" class="numchildren" onclick="return togglecomment(this)">(0 enfant)</a>
```

Figure 2: First HTML lines of a post in Reddit.

As shown in the Figure 2, the HTML is not as straight forward as JSON and contains a lot of information that are not relevant because not patient-generated data.

DATA RETRIEVAL

SELECTION CRITERIA

Selection criteria may vary according to the sources but in general (and this is true for Twitter and Reddit), they are very similar. Data search is done by keyword and more specifically a drug. For Twitter, the drug is searched in the posts and when it is found, a new record is created. For Reddit, the drug is searched in the title of posts and when it is found, a new record is created for every comment in the post. There is no limit in time or numbers. All data which are publicly available are taken. Data exclusion is performed after the extraction, during the cleaning and processing of the data.

The drugs we worked with have been determined according to the most in shared ones into PatientsLikeMe®, a health-related website. These drugs are:

- Aspirin
- Duloxetine
- Pregabalin
- Clonazepam
- Baclofen
- Levothyroxine
- Ibuprofen
- Gabapentin

TWITTER EXTRACTION

Twitter data are publicly available and easily accessible because Twitter provides an Application Programming Interface (API) which is a tool to help software application building. It allows to query tweets directly from Twitter according to some keywords, location or even language (see Figure 3 as an example).

To detect ADRs, several researches are made using the Twitter API using the chosen drugs as keywords. Only English tweets are targeted because the other source Reddit is mainly in English. Tweets are extracted daily as the Twitter API is restricted in time (tweets older than 7 weeks can't be retrieve) and in numbers (2800 tweets each 15 min). These limits can be removed by using the non-free version of the API called GNIP.

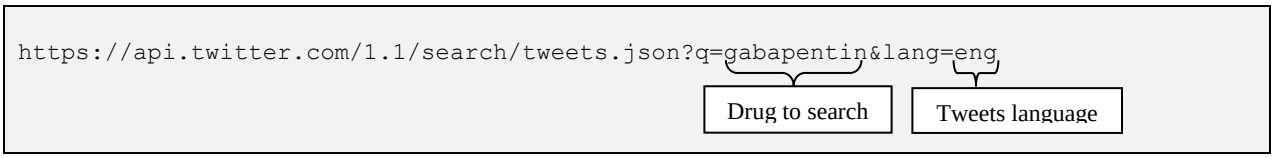

Figure 3: Example of the Twitter API usage.

The Twitter API is similar to a classic Uniform Resource Locator (URL) which can be used in a browser or with a programming language.

When the URL is used, a JSON file is returned containing tweets with the drug searched. The JSON file is not kept in its initial format as many data provided by Twitter are not relevant for the analysis (e.g. profile background color, profile image…).. Only the following data are kept:

- Tweet date creation
- If the tweet is a repost
- Text of the tweet
- The user id
- The drug researched

REDDIT EXTRACTION

Reddit data extraction is more complicated as HTML need to be analyzed to get all the valuable information. The first step is to observe the website to locate the relevant information was. When the overall structure is understood, a script can be written to download the webpages according to the drug to search.

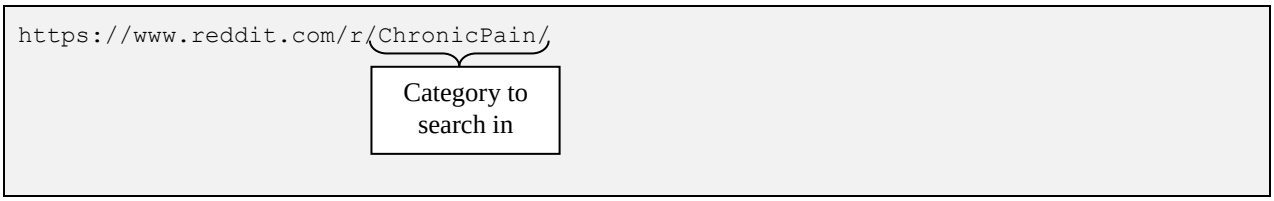

Figure 4: Example of URL to use to download a Reddit page with 'Gabapentin' posts

The figure 4 shows an example of URL which can be used to download the webpages about chronic pain where it is possible to find 'Gabapentin' posts from the Reddit website. This URL doesn't return a JSON file; but returns the source code of the page which is in HTML. This code contains a lot of unusable information which is here to delimitate some parts of the website, to structure it or related to the design. Markup structure embeds a lot of data making the automation more difficult to develop but this development has to be done only once. Indeed, drugs webpages on Reddit are identical in their structure allowing the script to be reused. When a certain drug is found in a post title, all the comments from this post are extracted. After an extraction from this source, several data can be kept:

- Comment date creation
- Text of the comment
- The user id
- The drug researched
- Tag of the comment
- Title of the post

DATABASE SPECIFICATION

Because the volume of data can be important, more than 3 000 records per day for Twitter, keeping them in flat file like Comma Separated Values is not recommended as it makes data selection difficult. Moreover, as web sources are used, their data format can change anytime and some new value which can be interesting to extract could be added. This is why a flexible storage is needed and a relational database is not suited.

It is to overcome the relational database limitation in terms of flexibility that MongoDB, a document-oriented database program, is used. This kind of database is designed for storing, retrieving and managing document-oriented information like semi-structured data[17]. It belongs to the NoSQL databases which provide a mechanism for storage and data retrieval modeled in another way than the tabular relations used in relational databases. The main goal is to have simplicity of design and more flexibility[18].

Date	Retweeted	Text	User_id	User_location	User_name	Drug
datetime.datetime(2017, 6, 19, 15, 57, 22)	False	"I took Gabapentin for my bach pain and now I have nausea…"	70470	"France"	"erwanlc"	"Gabapentin"

Figure 5: Example of relational storage

In Figure 5, a tweet is stored in a relational database. Twitter may decide to add a new value in its JSON file like the "user mental condition". If this information is relevant, a new column needs to be added to our table requiring a database modification. Document-oriented database doesn't need modification if a new field is added so the automation of extraction and storage are easier. In Figure 6, it is possible to see two documents stored in a MongoDB collection (the equivalent of the relational table) which don't share the same properties: Document 1 has a "retweeted" property and Document 2 has the "medical_condition".

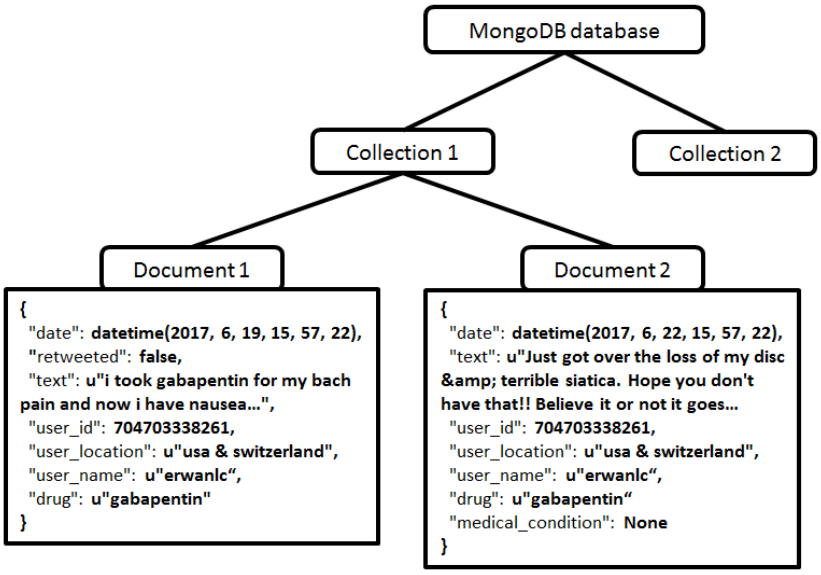

Figure 6: Example of MongoDB architecture and storage.

DATA IDENTIFICATION AND CODING

DATA CLEANING

After raw data storage, data cleaning step is needed before detection can be done. This step is very important to make data readable and usable. It consists of removing duplicates. This is especially done for the Twitter data: tweets reposted by another user don't carry new information and are not personal to the user who reposts. Then, unreadable words are removed as they mostly consist in emotes. Hyperlinks and punctuation are also deleted because no ADRs or drug can be detected in them.

This step allows reducing the size of the data processed, thus reducing execution time and allowing an easier manual verification (see Figure 7).

Raw text	`Just got over the loss of my disc &amp; terrible siatica. Hope you don't have that!! Believe it or not it goes… https://t.co/gZyAd5mxUB`
Cleaned text	`Just got over the loss of my disc terrible siatica Hope you dont have that Believe it or not it goes`

Figure 7: Example of a tweet before and after the cleaning

The cleaning process is identical for every source but differences between raw and cleaned text are not always visible depending on the consistency of the initial text.

ADR DETECTION AND CODING

For the detection, two dictionaries are used:

- The Medical Dictionary for Regulatory Activities (MedDRA®), a clinically validated international medical terminology dictionary[19].

- The WHO Drug Dictionary Enhanced (WHO-DDE®), an international classification of medicines created by the WHO Programme for International Drug Monitoring used for drug identification in spontaneous ADR reporting[20].

MedDRA is used to code ADRs in the text of the Twitter and Reddit posts. This dictionary was merged with the ADR San Diego lexicon which provides some expression used in social media as users don't always use medical expression[22]. For the drug detection and coding, it is the WHO-DDE which is used as provided by the Uppsala Monitoring Center.

For each record that is cleaned, all items from the MedDRA and the WHO-DDE are searched in the text. If any item is found, all related information regarding the item is added to the record.
The coding process is the following:

- Selection of the following variables: text for Twitter, text and tag list for Reddit.
- Search of the MedDRA and WHO-DDE items (word sequence included) in the variables.
- If an item is found as a substring of a variable (perfect match), the whole row of the dictionary is added to the post.

For ADRs detection, it is the variable LLT_NAME which is searched in the text of the posts. LLT_NAME is the Lowest Level Terms which is the most specific level. This variable reflects how an observation might be reported in practice[21]. When a LLT_NAME is found in a post, the whole row is returned.

For the drug detection, TSYN variable is searched. TSYN is the lowest level term for the drug. The search is made in the text of the records but also in the drug variables containing the drug researched which still require medical coding. A drug search is also performed in the text of the post to see if users are taking additional medication.

Regardless of the source, if no ADR is detected in a post, a manual revision should be performed as the following:

- Read the text variable
- If there is one, identify the ADR expression used
- Give the Preferred Term equivalence of the ADR

After revision, the identified expression is added to the ADR lexicon to automatically detect it if reused. Revision is not done for drug names because they are always present in the posts as they are the keywords for extraction.

Final result can be seen in Figure 8. Once the records are coded, they are stored in the database.

Condition	Drug	Cleaned_text	…
Peripheral Neuropathy	**Gabapentin**	Horrid drug Did little for PN and caused severe **swelling** in legs and diminished my cognitive abilities	…

Record

TPN	TSYN	TSYNC	…
GABAPENTIN	**GABAPENTIN**	1003001001	…

Drug detection result

LLT_CODE	LLT_NAME	PT_NAME	SOC_NAME	…
10042674	swelling	Swelling	General disorders and administration site conditions	…

ADR detection result

Figure 8: Example of drug and ADR detection on a Twitter post.

RESULTS

For Twitter, data were collected daily over a 3 months period, from April to July 2017 for a total of 23,695 records. Depending on the number of data extracted, the processing time from extraction to medical coding ranged from less than 1 to 5 minutes. Over all the data, potential ADRs were automatically found in 37% of the tweets.

For Reddit all the comments since June 2016 have been collected for a total of 145,770 records.

Drug	Twitter	
	Potential ADR (% # total)*	*# total*
Aspirin	3,736 *(41%)*	9,163
Duloxetine	126 *(38%)*	333
Pregabalin	298 *(39%)*	760
Clonazepam	175 *(18%)*	959
Baclofen	174 *(46%)*	375
Levothyroxine	139 *(41%)*	340
Ibuprofen	3,678 *(34%)*	10,695
Gabapentin	427 *(40%)*	1,070
Total	**8,753** ***(37%)***	**23,695**
Median	*236.5 (40%)*	*859.5*

*Considering that reporter is identified.

Figure 9: Detection result

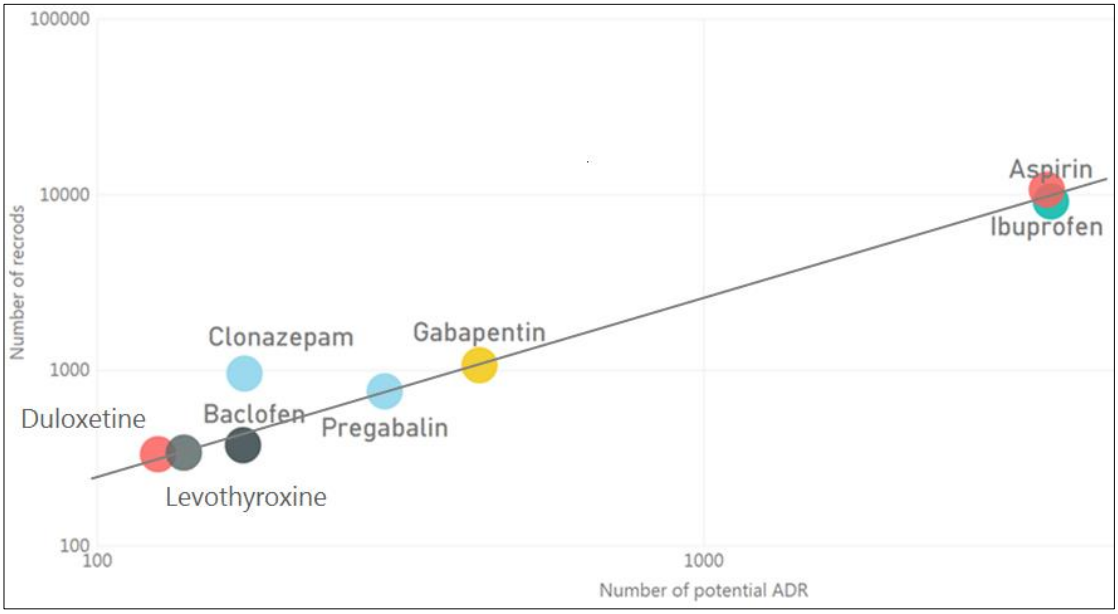

Figure 10: Potential ADR ratio on Twitter

The figure 9 and 10 shows the number of potential ADR detected on the Twitter data. Even if some drugs have few records, the median for the detection is still 40%. The drug with the lower rate of automatic detection is 'Clonazepam' with 18% of potential ADR. The drug with the highest rate of automatic detection is 'Baclofen' with 46% but it has one of the lowest numbers of tweets.

Drug	Twitter *# total (% of total drugs)*
Ibuprofen	10,695 *(45%)*
Aspirin	9,163 *(39%)*
Gabapentin	1,070 *(4.5%)*
Clonazepam	959 *(4.0%)*
Pregabalin	760 *(3.2%)*
Baclofen	375 *(1.5%)*
Levothyroxine	340 *(1.4%)*
Duloxetine	333 *(1.4%)*
Total	**23,695**
Median	*859.5*

Drug	Reddit *# observed (% of total drugs)*
Ibuprofen	85,586 *(58.7%)*
Aspirin	54,308 *(37.3%)*
Gabapentin	1,405 *(1.0%)*
Duloxetine	1,368 *(0.9%)*
Baclofen	1,279 *(0.9%)*
Clonazepam	815 *(0.6%)*
Pregabalin	565 *(0.4%)*
Levothyroxine	444 *(0.3%)*
Total	**145,770**
Median	*1,324*

Figure 11: Records extracted

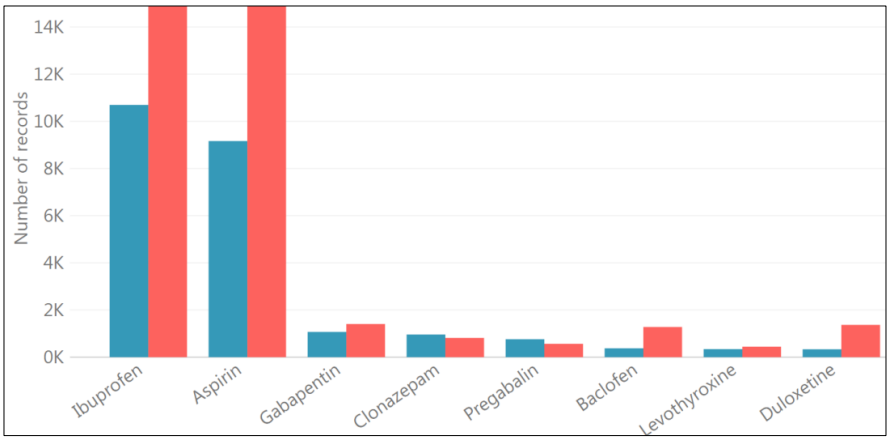

Figure 12: Drug representation in Twitter vs Reddit

The figure 11 and 12 shows the number of records extracted for Twitter versus the number of comments observed in Reddit. There are much more records for Reddit (6x more records than for Twitter), but more than 95% of the data stored concerns 'Ibuprofen' or 'Aspirin'. 'Aspirin' and 'Ibuprofen' are the most represented drugs for each source with more than 9,000 tweets for each on Twitter and more than 50,000 for each on Reddit.

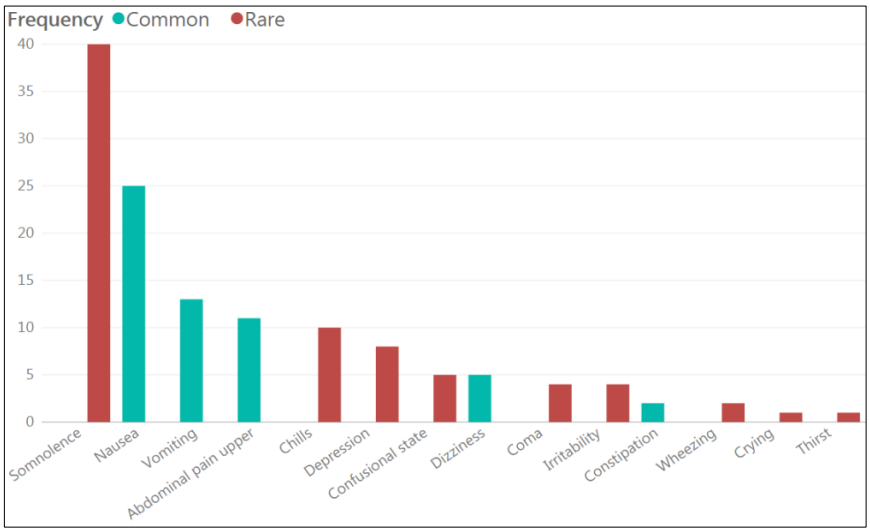

Figure 13: Ibuprofen (mixed strength), ADR detection vs literature

Figure 13 compare the number of real ADR detected in the tweets with the drug Ibuprofen to their known frequency in the literature and usage notice. The data repartition is not similar as we can see that the most common ADR in tweets is Somnolence which is supposed to be rare.

Continent	Ibuprofen	Aspirin	Clonazepam	Total
Asia	112 *(23.9%)*	348 *(74.2%)*	9 *(1.9%)*	469
Europe	1217 *(48.3%)*	1254 *(49.7%)*	50 *(2%)*	2521
Oceania	146 *(33.8%)*	279 *(64.6%)*	7 *(1.6%)*	432
Undetermined	2418 *(56.1%)*	1699 *(39.4%)*	193 *(4.5%)*	4310
North America	4517 *(51.4%)*	3942 *(44.9%)*	325 *(3.7%)*	8784
Middle East	21 *(24.4%)*	62 *(72.1%)*	3 *(3.5%)*	86
Afrique	60 *(25.8%)*	172 *(73.8%)*	1 *(0.4%)*	233
South America	59 *(23.8%)*	70 *(28.2%)*	119 *(48%)*	248
Total	**8550** ***(50.1%)***	**7826** ***(45.8%)***	**707** ***(4.1%)***	**17083**
Median	*129*	*313.5*	*29.5*	*450.5*

Figure 14: Ibuprofen, Aspirin and Clonazepam representation by continent

The figure 14 shows for each continent the number of tweets dealing with 'Ibuprofen', 'Aspirin' and 'Clonazepam'. A lot of data (~25%) don't have a determined location, but for the others, the repartition between continents is similar.

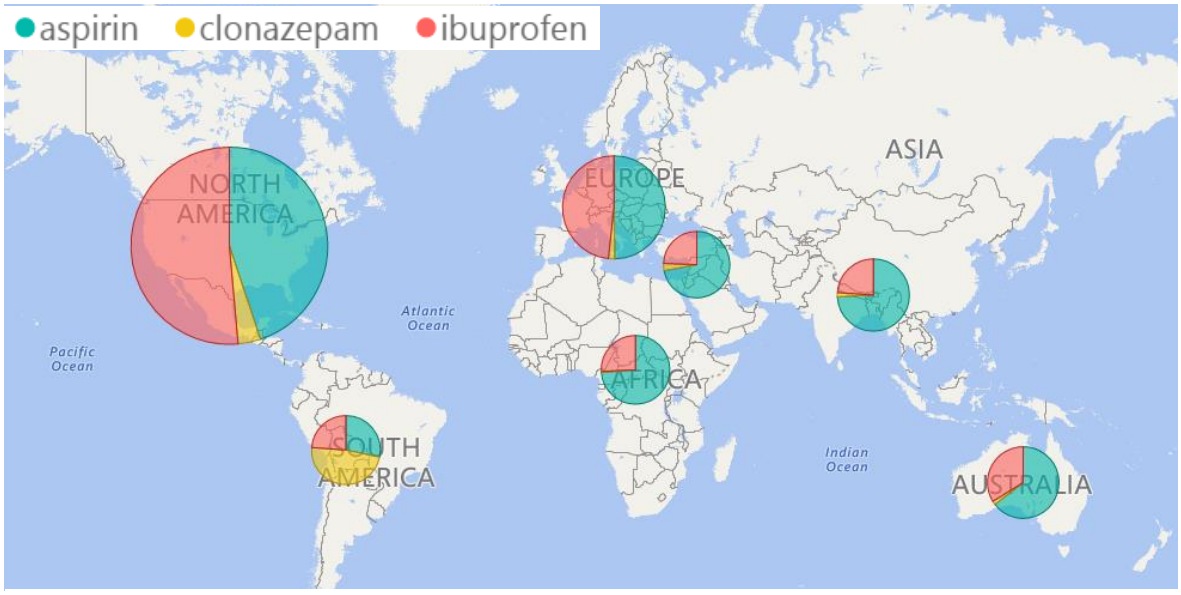

Figure 15: Map of the Ibuprofen, Aspirin and Clonazepam representation by continent

The figure 15 shows the result of figure 15 on a map. It allows to distinguish 3 groups of continents:

- North America and Europe
- Middle East, Africa, Australia and Asia
- South America

DISCUSSION

EXTRACTION

Automated ADR detection and storage in social media is possible. The method is successful both for Twitter and Reddit and should be also for other kinds of health-related websites with small coding additions in the extraction steps. The total number of data is increased compared with the use of one social media only, allowing better drugs representation but a lot of data coded by the ADR dictionary are not ADR but medical conditions. The overall results on the detection are similar to the previous study using lexicon-based methods. However, medical coding and storage is done in addition in our method to help further analysis and comparison between sources.

Figure 11 shows an example of a detection result: it appears that drug representation is not homogenous. Some drugs are well represented for both sources like Ibuprofen and Aspirin but more specialized drugs were found is much less records. The two sources are complementary and it demonstrates that using only one source is not enough. A third or a fourth source would probably add more value to drug with low representation like Levothyroxine. The best would be to have a list of sources with their main drugs representation. It would allow to better target the sources if some particular data are required.

In general, there is much more data for Reddit than for Twitter. It can be explained by the time limitation imposed by Twitter for its free version of the API. On Reddit such limitation doesn't exist and it is possible to extract data from the beginning of the website. However, Reddit data are probably more inconsistent as tweets because there is no limitation in the size of the text and the platform is mostly used by male (74% of users[13]) which is not representative of the worldwide population. Adding a health-related source like the website PatientsLikeMe (PLM) could add a lot of value but such sources often provide their data with a paying agreement. In return, the data are often more structured. Indeed, PLM users are registered on the website to share medical issues and information so they tend to be more precise in their posts: character numbers are not limited, drug dosage and side effects are present. Twitter and Reddit users often post their state of mind so they do it quickly with more elaborate expression which can be difficult to detect.

ADR DETECTION

Amongst the records without ADRs detected, it seems an important part of them doesn't have any mention of ADR. The other part seems to have ADRs but the expressions used are too complex for the detection (e.g.: "*SjogrensOrg Im male 65 yo Taking Plaquenil and Gabapentin Is feeling like I have to stretch all over all the time part of it*"). This barrier language can be even stronger with other languages with a more diverse medical vocabulary for which dictionaries don't always exist. As this information has not been quantified and results from visual observation it still needs real evidence.

These evidences are lacking because manual revision is needed to evaluate the record for potential valid ICSR. This is the main issue of the method. As shown in the Figure 9, a lot of records need manual revision which is very time consuming. To facilitate it, a solution to estimate the probability of ADR presence could be considered. The use of word embedding to produce a vector space from the corpus of posts, followed by clustering, could allow to separate words or assembly of words in a cluster. Then, classification could be done over the words or expressions to have a probability, weighted by the clustering result, of ADR presence. Posts with a probability of ADR presence higher than a certain level would be reviewed manually while the others would be considered to be without any ADR.

DEMOGRAPHIC DISCUSSION

In general, data from social media are not representative of a population as a whole. The elderly, for example, don't use the internet a lot. Data can be also very noisy because some accounts are only used for advertisements, promotion or robots programmed to post regularly. Thus, data can be stored if a drug is mentioned but they don't represent a real person with real problems. Analyzing such data can lead to misinformation and is a challenge as reporter must be identified as a real patient to classify data as ADRs

To get more information, the demography of the users should be completed such as the age, the sex, the ethnicity … It is very difficult to retrieve the information from Twitter and even more for Reddit. That's why to perform such analysis we would recommend to use social media like PLM.

To illustrate these difficulties, the Figure 15 presents the location of the 16,376 tweets about Ibuprofen and Aspirin. More than 25% of the tweets do not present information to identify the country / region or continent. And for the ones where the location could be identified, we do not have certitude on the information reliability.

Considering that the more a drug is used, more tweets exists, it is interesting to observe (Figure 14 & 15) the following:

- The use of 'Aspirin'/'Ibuprofen' ratio is ~75% / 25% for most of the world,
- For occidental regions, Ibuprofen is more used with a ratio 'Aspirin'/'Ibuprofen' around 50% / 50%

The exception is for South America where the most used drug is Clonazepam. No explanations have been found yet but it raises some questions which could be the subject of a specific paper. However, there is a high amount of data where location could not be determined and which is probably source of bias.

The data obtained here may not qualify yet for signal detection. A lot of them can be categorized as off label or misused and sometimes there is not enough information. Moreover, obtained data are not always ADR as users' symptoms not related to drugs are also coded. In order to do Pharmacovigilance, a second system of classification, which could be the same as the one presented above, may reduce the risk of getting non ADR value.

However, stored data still carry some valuable information. They allow to:

- Do some drug watch
- See the evolution of some drug/symptom combination used through time
- Detect unusual drug/symptom combination

CONCLUSION

Patient-generated health data in social media is a potential source for ADR reporting that are currently receiving significant attention in Pharmacovigilance field and public health protection measures. Different approaches to ADR data extraction from social media were proposed in literature. This method is structured for detection and extraction of ADRs from multi-source unstructured data collection system. It also retrieves demographic information which carries valuable information on drug usage around the world.

The provided method works well for known ADR terms but is difficult to use in production because of the time consuming manual revision needed to isolate real ADRs from symptoms or other medical events. A solution using new machine learning techniques to automatically estimate the probability of ADRs presence might be used. This might have a crucial influence and visible role on evaluating the medicines products' benefits versus potential risk and might qualify for safety signal detection. This will be the topic for further studies.

DISCLAIMER

The data contained in this paper was obtained for demonstration purposes only (from Twitter and Reddit) using the techniques presented. Anyone analysing data from social media sites, either public or membership based, should investigate the source terms and conditions relevant to data reuse and obtain any necessary permissions if required.

REFERENCES

1. Definition of SOCIAL MEDIA [Internet]. Merriam-webster.com. 2017 [cited 31 July 2017]. Available from: https://www.merriam-webster.com/dictionary/social%20media

2. Twitter [Internet]. En.wikipedia.org. 2017 [cited 25 April 2017]. Available from: https://en.wikipedia.org/wiki/Twitter

3. Press T, Press A. Number of active users at Facebook over the years [Internet]. Finance.yahoo.com. 2017 [cited 25 April 2017]. Available from: https://finance.yahoo.com/news/number-active-users-facebook-over-years-214600186--finance.html

4. Ginn R, Pimpalkhute P, Nikfarjam A, Pakti A, O'Connor K, et al. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark [Internet]. 2017 [cited 13 July 2017]. Available from: http://www.nactem.ac.uk/biotxtm2014/papers/Ginnetal.pdf

5. White R, Harpaz R, Shah N, DuMouchel W, Horvitz E. Toward Enhanced Pharmacovigilance Using Patient-Generated Data on the Internet [Internet]. 2017 [cited 13 July 2017]. Available from: http://onlinelibrary.wiley.com/doi/10.1038/clpt.2014.77/full

6. Guideline on good Pharmacovigilance practices (GVP) - Module VI [Internet]. 1st ed. 2017 [cited 25 April 2017]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf

7. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, L Smith K. Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions [Internet]. 2017 [cited 13 July 2017]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419871/

8. Nikfarjam A, Gonzalez G. Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments [Internet]. 2017 [cited 13 July 2017]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243273/

9. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features [Internet]. 2017 [cited 13 July 2017]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4457113/

10. Social media [Internet]. En.wikipedia.org. 2017 [cited 28 June 2017]. Available from: https://en.wikipedia.org/wiki/Social_media

11. Social media mining [Internet]. En.wikipedia.org. 2017 [cited 28 June 2017]. Available from: https://en.wikipedia.org/wiki/Social_media_mining

12. 30 Facts & Statistics On Social Media And Healthcare [Internet]. Patient Access, Referral Management & E-Consult Software - referraMD. 2017 [cited 28 June 2017]. Available from: https://getreferralmd.com/2017/01/30-facts-statistics-on-social-media-and-healthcare/

13. Reddit [Internet]. En.wikipedia.org. 2017 [cited 20 September 2017]. Available from: https://en.wikipedia.org/wiki/Reddit

14. Most famous social network sites worldwide as of April 2017 r. Global social media ranking 2017 | Statistic [Internet]. Statista. 2017 [cited 28 June 2017]. Available from: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

15. STATS | Twitter Company Statistics - Statistic Brain [Internet]. Statistic Brain. 2017 [cited 28 June 2017]. Available from: http://www.statisticbrain.com/twitter-statistics/

16. API Overview [Internet]. 2017 [cited 28 June 2017]. Available from: https://dev.twitter.com/overview/api

17. Document-oriented database [Internet]. En.wikipedia.org. 2017 [cited 30 June 2017]. Available from: https://en.wikipedia.org/wiki/Document-oriented_database

18. NoSQL [Internet]. En.wikipedia.org. 2017 [cited 30 June 2017]. Available from: https://en.wikipedia.org/wiki/NoSQL

19. MedDRA [Internet]. Meddra.org. 2017 [cited 30 June 2017]. Available from: https://www.meddra.org/

20. UMC | WHODrug Portfolio [Internet]. Who-umc.org. 2017 [cited 30 June 2017]. Available from: https://www.who-umc.org/whodrug/whodrug-portfolio/

21. MedDRA Hierarchy | MedDRA [Internet]. Meddra.org. 2017 [cited 5 July 2017]. Available from: https://www.meddra.org/how-to-use/basics/hierarchy

22. DIEGO Lab [Internet]. Diego.asu.edu. 2017 [cited 27 September 2017]. Available from: http://diego.asu.edu/index.php?downloads=yes

ACKNOWLEDGMENTS

This paper was made possible thanks to the Pharmacovigilance team of Keyrus Biopharma who helped us in each step of the project, the people who helped for the review and everyone who encouraged us and helped us in this work with their kind words and good mood.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Erwan Le Covec
Keyrus Biopharma
Chaussée de Louvain 88
Lasne / 1380
Belgium
Email: erwan.lecovec@keyrus.com

Essam Ghanem
Keyrus Biopharma
Chaussée de Louvain 88
Lasne / 1380
Belgium
Email: essam.ghanem@keyrus.com

Stéphane Chollet
Keyrus Biopharma
18/20 rue Clément Bayard
Levallois-Perret / 92300
France
Email: stephane.chollet@keyrus.com

Web : http://www.keyrusbiopharma.com