MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-322

Mining Adverse Drug Reactions in Social Media with Named Entity Recognition and Semantic Methods

Xiaoyi Chen[a], Myrtille Deldossi[b], Rim Aboukhamis[c], Carole Faviez[d], Badisse Dahamna[e,f,g], Pierre Karapetiantz[a], Armelle Guenegou-Arnoux[a], Yannick Girardeau[h,i], Sylvie Guillemin-Lanne[b], Agnès Lillo-Le-Louët[c], Nathalie Texier[d], Anita Burgun[a,h,i], Sandrine Katsahian[a,h,i]

[a] *INSERM, UMRS1138, équipe 22, Centre de Recherche des Cordeliers, Paris, France*
[b] *Expert System, 75012 Paris, France*
[c] *Centre Régional de Pharmacovigilance, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France*
[d] *Kappa Santé, 75002 Paris, France*
[e] *Service d'Informatique Biomédicale, CHU de Rouen, France*
[f] *LITIS-TIBS EA 4108, 76031 Rouen Cedex, France*
[g] *INSERM, U1142, LIMICS, 75006 Paris, France*
[h] *Université Paris Descartes, Sorbonne Paris Cité, UMRS1138, Centre de Recherche de Cordeliers, Paris, France*
[i] *Département d'Informatique Hospitalière, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France*

Abstract

Suspected adverse drug reactions (ADR) reported by patients through social media can be a complementary source to current pharmacovigilance systems. However, the performance of text mining tools applied to social media text data to discover ADRs needs to be evaluated. In this paper, we introduce the approach developed to mine ADR from French social media. A protocol of evaluation is highlighted, which includes a detailed sample size determination and evaluation corpus constitution. Our text mining approach provided very encouraging preliminary results with F-measures of 0.94 and 0.81 for recognition of drugs and symptoms respectively, and with F-measure of 0.70 for ADR detection. Therefore, this approach is promising for downstream pharmacovigilance analysis.

Keywords:

Social Media; Pharmacovigilance; Data Mining

Introduction

The rapid expansion of the Internet and social media is changing the way people gather information about disease and treatment, as well as how they share personal health experiences with others [1]. The *Digit in 2016* [2] reported that, in France, 86% of the population are active internet users. This proportion is higher than Western Europe's average (83%) and slightly lower than North America's average (88%). Various questionnaire statistics [3-5] showed that a large proportion of French people (46% to 71%) use the Internet to seek medical or health related information. Many people also use social media, such as forums, to communicate with others with the same health concerns and share information related to their illnesses, feelings, medication use and many other aspects [6], which offers promising opportunities for public health surveillance with a rich internet-based, patient-generated source.

The World Health Organization (WHO) defines Pharmacovigilance as "the science relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problems". It begins during clinical trials and continues after the drug is released onto the market. However a study [7] showed that 60% of potentially fatal ADRs were not described in initial drug labels and 39% were not included in any report of randomized controlled trials. The main pharmacovigilance tools are spontaneous reporting systems, driven by drug agencies, like the U.S. FDA' (Food and Drug Administrations) Adverse Event Reporting System (FAERS) which gathers voluntary reports by healthcare professionals and consumers (59% by professionals Vs. 41% by consumers in 2006, and 46% Vs. 54% in Q1 2015 [8]). It can also include Phase IV clinical trials driven by pharmaceutical companies and governmental agencies [9]. Despite such systems, the underreporting of ADRs by the patients as well as by the health professionals remains a significant limitation [10;11].

Several studies have already demonstrated the value of mining ADR from social media posts [12-14]. However, in contrast to the numerous studies in social media, the potential of utilizing this data for pharmacovigilance has not yet been fully exploited. It represents only 0.5% of publications with "social media" (SM) keyword query in the PubMed database (Figure 1A). Figure 1B shows that the number of publications with "SM + pharmacovigilance" as keywords has increased exponentially in the last five years.

A recent scoping review [11] outlined five complete steps that should be taken for processing ADR extraction in social media: (1) data collection, (2) preprocessing, (3) entity recognition (for drugs and symptoms), (4) identifying the relationship between drug and symptom, and (5) results evaluation. Since content and language of medical social media differ from those of general social media and of clinical documents, specific text mining methods or techniques based on Natural Language Processing (NLP) are necessary for step (3) and step (4) in order to identify medical concepts (such as drugs, symptoms, etc.) and relations among them [15]. It is evident that the performance of the text mining methods plays a decisive role in ADR signal detection.

From a text mining perspective, the key challenge is that Internet users' expressions are usually informal and colloquial, especially when they describe their feelings and symptoms. However, researches have progressed using (i) various data sources, such as forum messages [16;17], Twitter micro blogs

[18;19] and Yahoo Wellness Groups [12], in (ii) different languages, (iii) diverse approaches, such as Support Vector Machine (SVM) and Conditional Random Fields (CRF) [20], and (iv) different medical terminologies, making it difficult to compare their performances [9].

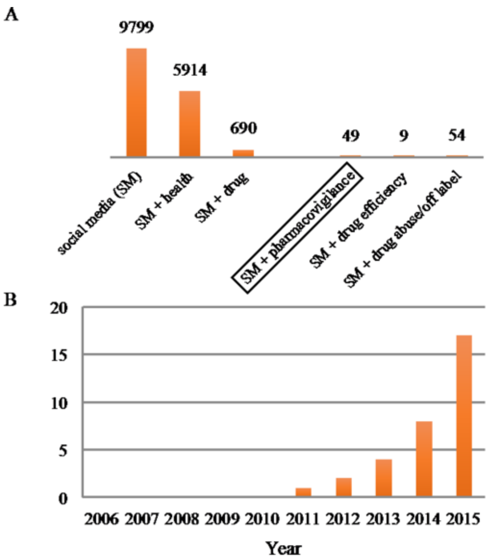

Figure 1 – (A) Number of publications with related key words (B) Distribution of "social media (SM) + pharmacovigilance" publications over 2006-2015

In this article, we introduce the approach developed in ADR-PRISM project in order to mine pharmacovigilance signals from social media. The utilized corpus, lexicons and methods of automated annotation are described in the next section. Moreover we highlight the evaluation protocol, which includes a detailed sample size determination and corpus constitution. The preliminary results are then discussed.

Materials and methods

Corpus

With the objective of extracting ADRs reported by patients on social media, we selected four French language, health related web sites:

- www.atoute.org
- forum.doctissimo.fr
- santemedecine.journaldesfemmes.com/forum/
- www.e-sante.fr/forums.

These sites were selected through search engines and the CISMeF web site, which is a catalog and index of French health resources on the Internet, and were evaluated using the Net scoring Tool [21].

An extractor targeting patients' messages using the HTML structure was then applied to extract messages from these forums. The extractor uses the names of drugs as key words to identify all discussions quoting one of the key words, then extracts and cleans the messages from the discussions (removing useless information, like ads, signatures and quotations). We selected 50 drugs of interest as input key words, and for each of these drugs, 20 discussions were randomly picked and extracted. This extraction conducted to the identification of 325 435 messages published between 2002 and 2014, corresponding to 967 distinct discussions.

The analysis showed that nearly 50% of the messages explicitly mentioned at least one symptom, and this ratio is in-between that of general forum posts (24%) [22] and that of drug reviews (80%) [23], which confirmed that our selection and evaluation of web sites and extraction were effective.

Lexicons

The thesaurus RacinePharma, which includes 5164 drug names, was used to identify drug mentions. This terminology is created by CISMeF and *Service d'Informatique Biomédicale du CHU de* Rouen (SIBM) and updated monthly according to the French *Base de Données Publique des Médicaments* (BDPM), which ensures that it covers all medications on the French market that might be mentioned on social media.

Medical Dictionary for Regulatory Activities (MedDRA) version 15.1 was used to identify medical terms including symptoms, signs, diseases, diagnoses, names and results of analysis etc. In this article, we will use "symptom" to refer to all these terms in order to facilitate writing. MedDRA has a five level structure with a classification in 26 medical disciplines (SOC – System Organ Class), then in HLGT (High Level Group of Terms), HLT (High Level Terms), PT (Preferred Terms), and finally LLT (Lowest Level Terms). The coding of ADRs is done by the LLT, which includes synonyms, lexical variants, sub-elements, familiar expression or "old" terms, thus suits for our study. On this basis, we put in place a strategy to automatically overcome orthographic variations or missed/added terms into the LLT, thus built an extended version of MedDRA in order to increase the coverage without predicting all lay vocabulary in social media. When assessing the performance of automatic recognition of symptoms, we decided to consider the PT level, which clusters the synonyms or lexical variants that might be used by different posters. For example, if we select the PT "anxiety" to evaluate, recognitions of all LLT under this PT, like "worry", "anxiety", "anguish", would be grouped and examined together. As MedDRA is a directed acyclic graph, there may exist multiple paths for the same entity. For example, the PT "scar" belongs to SOC "skin and subcutaneous tissue disorders" and also SOC "injury, poisoning and procedural complications". In such situations, all possible hierarchies would be considered in the same manner.

Annotation

The Smart Taxonomy Facilitator (STF) Skill Cartridge™ developed by Expert System was applied on the initial corpus. It combines a rule-based approach and a dictionary-based approach The latter includes two main technologies: (i), Fuzzy Term Matching, to take into account possible variants of the terms present in the taxonomy, thus reducing the number of false negatives; (ii) Relevance Scoring, which applies a series of heuristics that assigns a score to each extracted concept, and thus eliminates the least relevant concepts in order to reduce false positives. STF also exploits lexical labels (part-of-speech tagging) to address ambiguity issues.

We integrated in the Skill Cartridge the domain specific dictionaries (RacinePharma and MedDRA) and some intern rules established by our pharmacovigilance experts and text mining experts. Then the fuzzy matching parameters were adapted respectively for drug and symptom recognition.

ADR corresponds to a ternary relationship between (i) a patient and (ii) a symptom related with (iii) a drug through a causal relationship. We identified the linguistic patterns that corresponded to the five major semantic relations between these three entities: administration (take, test, try, treatment, intake of, etc.), causal relationship (cause, give, result of, since, because of, etc.), sensation (suffer, feel, etc.), drug stop (stop to avoid, to arrest, etc.) and intolerance (endure, allergy, etc.). With the pre-defined linguistic patterns, the ADR Skill Cartridge™ is able to identify multiple relationships between one or more drugs and/or symptoms within one sentence.

Evaluation

Protocol overview

ADR mining from social media may have two different utilisations: (i) routine signal detection for public health and surveillance; and (ii) focused drug- (or symptom-) signal detection, mainly for pharmaceutical industry. We established a protocol to evaluate the performance of the recognition of drugs, symptoms and their relationships in those two contexts. We will describe in the next sections (1) the constitution of sub-corpus, i.e. selection of drugs and symptoms of interest and determination of sample size, (2) the establishment of gold standard, i.e. manual annotation and the guideline of manual annotation, (3) the statistics analysis for comparison.

Data sets

We expected to evaluate the general performance on all identified drug and symptom mentions, and also the performance on certain specific concepts. We therefore selected 12 drugs, including the most frequent in the corpus, the most sold in France in 2013, the most interesting according to the pharmacovigilance experts and we selected some drugs randomly. We selected 9 symptoms, based on similar principles, i.e., the most frequent, the most interesting and some randomly picked ones. For each selected concept, the sample size is calculated under the hypothesis of precision (or recall) = 0.5 ± 0.15, with a significant level of 0.05. The posts containing at least one of the selected concepts were then pooled to build the sub-corpus for evaluation.

Manual annotation

An annotation guideline was established for human experts to annotate all words that refer to a drug or a symptom, and then annotate for every pair of drug and symptom whether there was a causal relationship between them in the context, without using any expert knowledge, experience or intuition to prejudge. Two pharmacovigilance experts with experience in ADR reporting, annotated blindly and independently a part of the sub-corpus. Both experts annotated a common part of the messages, which aims to estimate the inter-annotator agreement, to perfect the guidelines and to improve the quality of manual annotation. We then considered this manual annotation as Gold Standard.

Comparison

The basic metrics used to evaluate the performance are precision, recall and their harmonic mean (F-measure). Three different result types are examined: false negative (FN) for non recognition of relevant terms, false positives (FP) for irrelevant positive recognitions and true positive (TP) for correct positive recognitions. The precision, recall and F-measure are defined respectively as Eq.1,

$$p = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \quad r = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \quad F_1 = \frac{2pr}{p+r}. \tag{1}$$

These statistics were computed globally and also specifically for each selected concept.

Format and implementation

The extracted forum messages were transformed in XML format for applying the STF Skill Cartridge. Automated annotations in XML format were parsed with R 3.3.1 *xml2* packages. The sampling for sub-corpus constitution was carried out by building the list of message IDs with R and then integrating the corpus in the Skill Cartridge.

Results

Description of dataset

With the corpus described above,

- 55 777 entities of drug names from 34 265 messages have been annotated by the Skill Cartridge with thesaurus RacinePharma, which concern 1383 distinct drugs;
- 429 424 entities of symptoms from 153 995 messages have been annotated by the Skill Cartridge with thesaurus MedDRA, which concern 4861 distinct MedDRA terms.
- On the basis of drug and symptom recognitions, 1385 ADRs have been identified from 1129 messages.

Table 1 shows an overview of our dataset.

Table 1 – Dataset overview

Corpus	Mentions			Discussion	Messages	Messages containing			
	Drugs	Symptoms	ADRs			Drugs	Symptoms	Both	ADRs
Total	55777	424924	1385	967	325435 % of line total	34265 10,5%	153995 47,3%	27394 8,4%	1129 0,3%
Atoute	5457	36314	139	261	27234	2825 10,4%	11421 41,9%	2229 8,2%	109 0,4%
Doctissimo	15980	90876	432	565	73266	9134 12,5%	32577 44,5%	7122 9,7%	358 0,5%
E-sante	34221	297010	806	108	224946	22232 10,5%	109792 48,8%	17978 8,0%	656 0,3%
Sante-medecine	119	724	8	33	286	74 10,5%	205 71,7%	65 22,7%	6 2,1%

The 12 drugs used for evaluation included four categories:

1. from the most frequent drugs in the corpus, PUREGON (follitropin beta), SPASFON (phloroglucinol, trimethylphloroglucinol), and TARCEVA (erlotinib);
2. from the top 15 most sold in France, ASPIRINE (acetylsalicylic acid), LEVOTHYROX (levothyroxin), and DOLIPRANE (paracetamol/acetaminophen);
3. from the most interesting according to the pharmacovigilance experts: METHADONE, DIANE 35 (ethinylestradiol, cyproterone acetate) and PROZAC (fluoxetin);

4. three drugs chosen at random: IXEL (milnacipran), CHAMPIX (varenicline) and GLIVEC (imatinib).

For symptoms, the 3 most frequent PT selected were “anxiety”, “pain” and “fatigue”; the 3 of interest PT were “death”, “hypersensitivity” and “injury”; the 3 randomly selected PT were “basedow’s disease”, “moderate mental retardation” and “fungal infection”.

Gold standard

For drug and symptom recognition, the sample size for gold standard was determined with the method detailed in the previous section, which corresponds to 45 messages per concept. As one message can contain several occurrences of the same concept, and also mentions of other concepts, the sub-corpus for evaluation carries more entities than 45 per concept. Then for ADR detection, we evaluated all 1129 messages containing at least one ADR identified by the automatic approach. Two human experts read all messages in the sub-corpus and annotated the drugs, symptoms, and the relationships between them as being an ADR or not. The gold standard corpus, corresponding to 785 mentions of drugs and 908 mentions of symptoms, was then used for comparison.

Comparison results

The sub-corpus for evaluating drug name recognition corresponds to 561 messages, in which the Skill Cartridge identified 721 occurrences of drugs corresponding to 27 distinct drugs. Table 2 shows the global scores and the scores obtained for each drug in terms of precision, recall, and F-measure.

Table 2 – Evaluation results of drug name recognition

Drug	Precision	Recall	F-measure
USE CASE			
PUREGON	1,00	1,00	1,00
SPASFON	1,00	0,98	0,99
TARCEVA	1,00	1,00	1,00
ASPIRINE	1,00	1,00	1,00
LEVOTHYROX	0,92	1,00	0,96
DOLIPRANE	1,00	1,00	1,00
METHADONE AP-HP	0,95	1,00	0,98
DIANE	1,00	0,92	0,96
PROZAC	1,00	0,99	0,99
IXEL	1,00	1,00	1,00
CHAMPIX	1,00	0,92	0,96
GLIVEC	1,00	1,00	1,00
…	…	…	…
OVERALL	**0,98**	**0,90**	**0,94**

The sub-corpus for evaluating symptom recognition corresponds to 401 messages, in which the Skill Cartridge identified 640 mentions concerning 59 distinct PTs. The results are shown in Table 3. Although, we present in this table only the PT level the MedDRA hierarchy makes it possible to display similar results, at other levels (SOC, HLGT, HLT).

Table 3 – Evaluation results of symptom recognition

MedDRA term (PT)	Precision	Recall	F-measure
USE CASE			
anxiety	0,98	0,86	0,92
pain	0,97	0,89	0,93
fatigue	0,98	0,97	0,98
death	0,99	0,97	0,98
hypersensitivity	1,00	0,92	0,96
Injury	0,92	0,96	0,94
basedow's disease	0,75	1,00	0,86
moderate mental retardation	1,00	1,00	1,00
Fungal infection	1,00	0,91	0,95
…	…	…	…
OVERALL	**0,98**	**0,69**	**0,81**

For ADR identification, we evaluated all messages containing ADR annotation and obtained a precision 0.78, recall 0.63 and F-measure 0.70. Taking into account the great challenges of the processing of social media texts, this relation detection result is encouraging and promising for downstream analysis.

Discussion

In this paper, we have described a methodology by which text messages on social media can be effectively transformed into a usable format for pharmacovigilance. The protocol of evaluation, which includes a detailed sample size determination, has been highlighted, which is often obscure in previously published works. We have obtained a nearly perfect accuracy on recognition of drug names, and good performance on recognition of symptoms and ADR relations. It seems that some types of mentions or relation patterns are easier to extract than others. One of the key issues is still the informal narrative in social media containing many grammatical errors, abbreviations, spelling mistakes and lay terminology.

Our performance of recognition (F-measure 0.94 for drugs and 0.81 for symptoms) is comparable with other studies in the domain. In CHEMDNER BioCreative IV challenge [24], the chemical compound and drug name recognition task reached an F-measure of 0.88, while the disease named entity recognition reached an F-measure of 0.86 in BioCreative V challenge [25; 26]. With Electronic Health Records (EHRs) data, the F-measure of drug name recognition varies from 0.73 to 0.89 [27]. With social media, although various studies have attempted to adapt different methods to this specific text data source, there is still a gap on recognition performance due to informal and colloquial expressions. Most pilot studies of mining ADRs from social media [28-30] have investigated for English language, and the F-measure ranged from 0.58 to 0.82. The performance depends mainly on the size and quality of dataset. A study of detecting drug effects from a Spanish health forum has obtained a precision of 0.48 and recall of 0.59 [31]. In French language social media, a study of automatic identification of drug-related medical conditions on drug review [23] obtained a F-measure of 0.95 for chemicals, 0.86 for signs/symptoms and 0.82 for diseases, however the relations are not considered in this work. Moreover their corpus and evaluation set are much smaller than ours.

Even if the automated annotation of ADR relations is now restricted to the co-occurrence of drugs and symptoms in the same sentence, our human annotators were asked to annotate all ADRs in the post regardless of the sentence boundary, which allows us to further assess the impact of the sentence restriction and eventually improve the performance of detection of relations across sentence boundaries. The method of evaluation presented in this article contains potentially a bias of overestimation of recall. The false negatives are actually underestimated due to the fact that we worked with messages containing at least one annotation for one of the selected drugs, symptoms, or one annotation of ADR. Rational for choosing this approach is that half of the messages did not exhibit any entity of interest (neither drug nor symptom).

Next step will be applying signal detection methods within the pharmacovigilance database issued from French social media. A comparison of these potential ADR signals with those detected from traditional reporting data will be performed. More work remains to examine how social media data can be incorporated into overall pharmacovigilance systems.

Conclusion

Our approach provides very encouraging preliminary results of recognition of drug names, symptoms, and ADRs in social media texts, which offer a promising basis for downstream analysis of routine or specific ADR signal detection.

Acknowledgements

This work was labeled by the competitiveness cluster Cap Digital and funded by the DGE and territorial collectivities (Ile de France and Haute Normandie) under the 16th FUI (Fonds Unique Interministériel) request for proposal through the ADR-PRISM project.

References

[1] R. Harpaz *et al.*, “Text mining for adverse drug events: the promise, challenges, and state of the art,” *Drug Saf.*, vol. 37, no. 10, pp. 777–790, Oct. 2014.

[2] “Digital in 2016,” *We Are Social Singapore*, 2016.

[3] H. Romeyer, “La santé en ligne,” *Commun. Inf. Médias Théories Prat.*, no. Vol. 30/1, Feb. 2012.

[4] Conseil National de l’Ordre des Médecins, “Vers une meilleure intégration d’Internet à la relation médecins-patients,” 2010. [Online]. Available: https://www.conseil-national.medecin.fr/article/vers-une-meilleure-integration-d%E2%80%99internet-la-relation-medecins-patients-982. [Accessed: 19-Apr-2016].

[5] V. Gombaut, “Deux ménages sur trois disposent d’internet chez eux,” *Insee Première*, Mar-2011. [Online]. Available: http://www.insee.fr/fr/themes/document.asp?ref_id=ip1340. [Accessed: 19-Apr-2016].

[6] S. Katsahian *et al.*, “Evaluation of Internet Social Networks using Net scoring Tool: A Case Study in Adverse Drug Reaction Mining,” *Stud. Health Technol. Inform.*, vol. 210, pp. 526–530, 2015.

[7] B. Seruga, L. Sterling, L. Wang, and I. F. Tannock, “Reporting of Serious Adverse Drug Reactions of Targeted Anticancer Agents in Pivotal Phase III Clinical Trials,” *J. Clin. Oncol.*, vol. 29, no. 2, pp. 174–185, Jan. 2011.

[8] “FAERS Reporting by Healthcare Providers and Consumers by Year,” *FDA Adverse Events Reporting System (FAERS)*, 24-Nov-2015. [Online]. Available: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070456.htm. [Accessed: 16-Dec-2016].

[9] A. Sarker *et al.*, “Utilizing social media data for pharmacovigilance: A review,” *J. Biomed. Inform.*, vol. 54, pp. 202–212, Apr. 2015.

[10] P. Tubert, B. Bégaud, J. C. Péré, F. Haramburu, and J. Lellouch, “Power and weakness of spontaneous reporting: a probabilistic approach,” *J. Clin. Epidemiol.*, vol. 45, no. 3, pp. 283–286, Mar. 1992.

[11] J. Lardon *et al.*, “Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review,” *J. Med. Internet Res.*, vol. 17, no. 7, Jul. 2015.

[12] B. W. Chee, R. Berlin, and B. Schatz, “Predicting Adverse Drug Events from Personal Health Messages,” *AMIA. Annu. Symp. Proc.*, vol. 2011, pp. 217–226, 2011.

[13] A. Yates and N. Goharian, “ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites,” in *Advances in Information Retrieval*, P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds. Springer Berlin Heidelberg, 2013, pp. 816–819.

[14] M. Abou Taam *et al.*, “Analysis of patients’ narratives posted on social media websites on benfluorex’s (Mediator®) withdrawal in France,” *J. Clin. Pharm. Ther.*, vol. 39, no. 1, pp. 53–55, Feb. 2014.

[15] K. Denecke, “Extracting medical Concepts from medical social media with clinical NLP tools: a qualitative study,” *Proc. Fourth Workshop Build. Evaluaion Resour. Health Biomed. Text Process.*, 2014.

[16] A. Yates, N. Goharian, and O. Frieder, “Extracting Adverse Drug Reactions from Forum Posts and Linking them to Drugs,” *Proceeding 2013 ACM SIGIR Workshop Health Search Discov.*, Aug. 2013.

[17] A. Benton *et al.*, “Identifying potential adverse effects using the web: a new approach to medical hypothesis generation,” *J. Biomed. Inform.*, vol. 44, no. 6, pp. 989–996, Dec. 2011.

[18] J. Bian, U. Topaloglu, and F. Yu, “Towards Large-scale Twitter Mining for Drug-related Adverse Events,” in *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, New York, NY, USA, 2012, pp. 25–32.

[19] K. Jiang and Y. Zheng, “Mining Twitter Data for Potential Drug Effects,” in *Advanced Data Mining and Applications*, H. Motoda, Z. Wu, L. Cao, O. Zaiane, M. Yao, and W. Wang, Eds. Springer Berlin Heidelberg, 2013, pp. 434–443.

[20] J. Lafferty, A. Mccallum, and F. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. 2001.

[21] “NetScoring : critères de qualité de l’information de santé sur l’Internet.” [Online]. Available: http://www.chu-rouen.fr/netscoring/. [Accessed: 22-Apr-2016].

[22] K. O’Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez, “Pharmacovigilance on twitter? Mining tweets for adverse drug reactions,” *AMIA Annu. Symp. Proc.*, vol. 2014, pp. 924–933, 2014.

[23] M.-H. François, G. Cyril, and Z. Pierre, “Identification of Drug-Related Medical Conditions in Social Media,” 2016.

[24] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, and A. Valencia, “CHEMDNER: The drugs and chemical names extraction challenge,” *J. Cheminformatics*, vol. 7, no. Suppl 1, p. S1, Jan. 2015.

[25] C.-H. Wei *et al.*, “Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task,” *Database J. Biol. Databases Curation*, vol. 2016, Mar. 2016.

[26] Q. Wei, T. Chen, R. Xu, Y. He, and L. Gui, “Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks,” *Database J. Biol. Databases Curation*, vol. 2016, 2016.

[27] S. Liu, B. Tang, Q. Chen, and X. Wang, “Drug Name Recognition: Approaches and Resources,” *Information*, vol. 6, no. 4, pp. 790–810, Nov. 2015.

[28] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez, “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*, Stroudsburg, PA, USA, 2010, pp. 117–125.

[29] H. Sampathkumar, X. Chen, and B. Luo, “Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model,” *BMC Med. Inform. Decis. Mak.*, vol. 14, p. 91, 2014.

[30] A. Nikfarjam, A. Sarker, K. O’Connor, R. Ginn, and G. Gonzalez, “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 22, no. 3, pp. 671–681, May 2015.

[31] I. Segura-Bedmar, P. Martínez, R. Revert, and J. Moreno-Schneider, “Exploring Spanish health social media for detecting drug effects,” *BMC Med. Inform. Decis. Mak.*, vol. 15, no. 2, pp. 1–9, 2015.

Address for correspondence

Xiaoyi Chen

Inserm UMRS 1138 - Equipe 22

15 Rue de l'école de médecine 75006 Paris, France

Email: xiaoyi.chen@inserm.fr