Data Mining and Knowledge Discovery

Automatic Discovery of Adverse Reactions through Chinese Social Media

--Manuscript Draft--

Manuscript Number:	DAMI-D-17-00435R1
Full Title:	Automatic Discovery of Adverse Reactions through Chinese Social Media
Article Type:	Manuscript
Keywords:	adverse drug reaction; Chinese social media; Natural Language Processing
Corresponding Author:	Mengxue Zhang Shanghai Jiao Tong University Shanghai, CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Shanghai Jiao Tong University
Corresponding Author's Secondary Institution:	
First Author:	Mengxue Zhang
First Author Secondary Information:	
Order of Authors:	Mengxue Zhang
	Meizhuo Zhang
	Chen Ge
	Quanyang Liu
	Jiemin Wang
	Jia Wei
	Kenny Zhu
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Despite tremendous efforts made before the release of every drug, some adverse drug reactions (ADRs) may go undetected and thus, cause harm to both the users and to the pharmaceutical companies. One plausible venue to collect evidence of such ADRs is online social media, where patients and doctors discuss medical conditions and their treatments. There is substantial previous research on ADRs extraction from English online forums. However, very limited research was done on Chinese data. In this paper, we try to use the posts from two popular Chinese social media as the original dataset. We propose a semi-supervised learning framework that detects mentions of medications and colloquial ADR terms and extracts lexicon-syntactic features from natural language text to recognize positive associations between drug use and ADRs. The key contribution is an automatic label generation algorithm, which requires very little manual annotation. This bootstrapping algorithm could also be further applied on English data. The research results indicate that our algorithm outperforms the hidden Markov model(HMM) and conditional random fields(CRF). With this approach, we discovered a large number of side effects for a variety of popular medicines in real world scenarios.

Dear reviewers,

Thank you so much for the precious comments. Below we itemize our responses and the revisions we have undertaken.

Reviewer #1

1.1 We have removed the references to Weibo in the abstract and introduction and only mention our efforts on Weibo and its infeasibility in Sec 2.2 data sources.

1.2 The number 46 was a typo. It should be 79 and we have modified it as such.

1.3 The wordings in the first column of Table 3 may be misleading. What we meant was "All features minus verbs before drugs (feature 1)", etc. What we wanted to show is how much F1/accuracy drops if one of the features is turned off. We have rephrased column 1 to be "Without feature 1", "Without feature 2", etc. Moreover, in Sec 2.3.2, we reformatted the list of features in a table for better readibility.

1.4 We put the algortithm in an algorithm environment and combined Sec 2.3.3 and 2.3.4 into one subsection with a section title "Automatic labeling by bootstrapping."

1.5 We plot an additional line in Fig 3, indicating the changes to training data size over the iterations.

1.6 We have added the results for pattern-based, HMM and CRF on auto labeled data and reorganized Table 4 accordingly.

Reviewer #2

2.1 What we actually meant was spontaneous reports can only be submitted by medical practitioners, and not normal patients, hence the data obtained through these reports represents just one source of information. Whereas information from medical forums which is studied by this paper comes directly from patients and may be more diverse and comprehensive, thus has the potential to cover rare ADRs. We have modified the wordings in this part of the introduction section.

2.2 In Sec 2.1.2, we have stated that we use the lexicon from Sougou to cover colloquial terms. We have also added some examples from Sougou to Sec 2.1.2 in the revision.

2.3 Table 5 actually shows the percentage increase in the number of sentences after homophone
transform and NOT ADR extension. We have revised the caption of that table accordingly.
Moreover, we have evaluated the precision/recall of our extended ADR lexicon and included this
new result in Sec 3.4.

2.4 Please see response 1.1.

2.5 The threshold 55 is not intended to determine if an ADR results from taking a drug.
Rather, if a drug name and an ADR are more than 55 words apart in the text, we
simply do not consider this pair at all. If the pair is within the 55-word distance,
whether it is an ADR relation is determined by our classifier.

2.6 Our initial training data consists of 300 positive and 300 negative samples (the rest of the
labeled data being tuning and test tests, see Sec 2.2). This is clearly not big enough to train a comprehensive classifier for all possible ADRs. Therefore we seek to
automatically enlarge the training data set in a bootstrapping fashion. The process of
automatic labeling is indeed a form of active learning, in which the package inserts provides
feedback in the learning process.

2.7 It is actually possible to have negative ranking scores. We have considered using the
ratio between positive and negative evidences, but later we decided that it serves the same
purpose.

2.8 The major contribution of this paper is not only using the Chinese social media data,
but also the bootstrapping framework to automatically enlarge training data. We believe
this framework can also be applied to English data and benefit that part of the world as well.

Reviewer #3
3.1 It is a common practice to train a binary classifier with balanced data, because i) the
distribution of the labels is often not known a priori; ii) we would have obtained a
biased classifier if the data is not balanced.

3.2 The original training data (300+300) was labeled by human and contains ADR pairs
that cannot be found from package inserts. Subsequently, it is true that we use the ADR or

indication information from package inserts to help us decide if an evidence is positive or negative when we automatically accumulate more training data. So indeed, during test time, if a pair of drug-ADR comes from a sentences with features similar to an indication relation, our classifier is inclined to predict false. Our classifier is not particularly trained to consider the "third class" which is no relation between the drug and the ADR, though such example may exist in the original negative training data. Having said that, we want to stress that our test data has been labeled by human who would label the "third class" as false. Therefore, there's nothing wrong with our evaluation results. Furthermore, in the revised version, we modify Fig 1 to make our overall framework more comprehensible. We also change the title of Sec 2.4 to "Baseline classifier techniques"

3.3 We have modified Table 4, to include more results for auto-labeled data (see response 1.6).

3.5 The features we used in this paper have previously been used one way or the other in previous relation extraction/classification work. We use verbs and prepositions as our features here because they collectively form dependency relations such as dobj and pobj, which provide important signals when it comes to the distinction of relations or predicate classification.

3.6 Please see response 1.1. We also changed the caption of Table 2 to "Catagory of drugs studied" and the term "Disease" to "Catagory" in the table header.

Noname manuscript No.
(will be inserted by the editor)

Automatic Discovery of Adverse Reactions through Chinese Social Media

Mengxue Zhang* · **Meizhuo Zhang*** · **Chen Ge** · **Quanyang Liu** · **Jiemin Wang** · **Jia Wei**** · **Kenny Q. Zhu****

Received: date / Accepted: date

Abstract Despite tremendous efforts made before the release of every drug, some adverse drug reactions (ADRs) may go undetected and thus, cause harm to both the users and to the pharmaceutical companies. One plausible venue to collect evidence of such ADRs is online social media, where patients and doctors discuss medical conditions and their treatments. There is substantial previous research on ADRs extraction from English online forums. However, very limited research was done on Chinese data. In this paper, we try to use the posts from two popular Chinese social media as the original dataset. We propose a semi-supervised learning framework that detects mentions of medications and colloquial ADR terms and extracts lexicon-syntactic features from natural language text to recognize positive associations between drug use and ADRs. The key contribution is an automatic label generation algorithm, which requires very little manual annotation. This bootstrapping algorithm could also be further applied on English data. The research results indicate that our algorithm outperforms the hidden Markov model(HMM) and conditional random fields(CRF). With this approach, we discovered a large number of side effects for a variety of popular medicines in real world scenarios.

Keywords adverse drug reaction · Chinese social media · natural language processing

*The authors contributed equally to this work. **Corresponding authors

Kenny Q. Zhu
Dept. CSE, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
E-mail: kzhu@cs.sjtu.edu.cn

Jia Wei
R&D Information, AstraZeneca, 199 Liangjing Road, Pudong, Shanghai, 201203, China
E-mail: Jenny.Wei@astrazeneca.com

1 Introduction

Determination of adverse drug reactions (ADR) is an important part of pharmaceutical research and drug development. Pre-marketing clinical trials are limited by the number of participants, the length of the study and the underlying economic burden for both the pharmaceutical companies and the patients. Some of the new adverse reactions to a drug are learned only when the drug is used in a wide spectrum of patients, with varied ethnicity, underlying diseases and a range of concomitant medication, in a post-launch setting. Furthermore, some reactions take a long time to develop a process which goes well beyond the pre-marketing development cycles of the drugs. For example, Vioxx, developed by Merck & Co, was approved by the FDA in May 1999 as a nonsteroidal anti-inflammatory drug to treat osteoarthritis, acute pain and dysmenorrhea. However, other Merck & Co sponsored studies, which were concluded or commenced after the drug was launched, indicated that it was associated with elevated risk of cardiovascular complications (Bombardier et al, 2000; Bresalier et al, 2005). In September of 2004, Merck withdrew Vioxx from the market because of concerns about increased risk of heart attack and stroke associated with long-term, high-dosage use. An FDA study estimated that Vioxx could have caused up to 140, 000 cases of serious heart disease in the US since 1999 (Graham et al, 2005). Regulatory authorities and pharmaceutical companies make tremendous effort in avoiding such incidences by conducting post-launch Phase IV clinical trials. In the United States, drug companies spend up to $12,000 per patient in Phase IV clinical trials, with an average of $5,856 [1]. Conducting such studies in an "*in silico*" fashion, i.e., collecting ADRs from pre-existing data sources, has become a valid complement, if not an attractive alternative, to costly Phase IV studies.

Recent years saw a growing research interest in mining adverse drug reactions from various data sources. Data sources can be divided into structured data and unstructured text data, and the approaches differ. Structured data primarily includes official adverse event reports collected by health authorities (Harpaz et al, 2010, 2012; Hahn et al, 2012; Gurulingappa et al, 2013). These reports are relatively easy to process due to their strict conformance to the adverse event reporting standards. However, these spontaneous reports can only be submitted by medical practitioners, and not normal patients. Hence, the data obtained through these reports represents just one source of information and it cannot catch many infrequent ADRs. Unstructured data so far includes biomedical literature, clinical notes or medical records, and online health discussions. These data sources pose more processing challenges because signals are embedded in natural language, which is inherently ambiguous and noisy. Biomedical literatures such as scientific papers are comparatively easier to mine (Wang et al, 2011; Yang et al, 2012a) since the medication and adverse reaction are referred to by their formal names. However, the information therein is not up-to-date and is sometimes biased. Clinical resources were tar-

[1] https://www.cuttingedgeinfo.com/2011/us-phase-iv-budgets/

geted using various methods, such as text mining for identifying ADRs from medicine uses (Warrer et al, 2012), rule-based methods to extract side effects from clinical narratives (Sohn et al, 2011) and retrospective medication orders along with inpatient laboratory results to identify ADRs (Liu and Chen, 2013). Privacy concerns and access restrictions are the biggest obstacles for its wide adoption. Compared to the above data sources, online social media, especially health discussion forums, provide the most comprehensive and timely information about medication use experiences. The large volume, colloquial use of natural language, spelling and grammatical errors are some of the major challenges in mining ADRs from such data sources.

Existing methods for social media text mining can be categorized into lexicon-based methods, statistical methods, rule-based method, advanced NLP and neural network. Most prior studies (Leaman et al, 2010; Yang et al, 2012b; Benton et al, 2011; Wu et al, 2013; Yates and Goharian, 2013; Liu et al, 2014; Jiang et al, 2013; Freifeld et al, 2014; Yeleswarapu et al, 2014) focused on expanding lexicons to find ADRs in text. In these lexicon-based methods, due to the novel adverse reaction phrases on websites, they could not recognize non-regular ADRs that are not contained in the lexicon. Besides, they suffer from poor approximate string matching caused by misspelled words. Some researchers instead utilized statistical (Li, 2011; Wu et al, 2012; Liu and Chen, 2013), rule (pattern) based methods (Nikfarjam and Gonzalez, 2011; Benton et al, 2011; Karimi et al, 2011; Yang et al, 2012b); When it comes to NLP techniques, common approaches used Support Vector Machine(SVM) and Conditional Random Field(CRF) to detect ADR from social media(Sharif et al, 2014; Sarker and Gonzalez, 2015; Jonnagaddala et al, 2016; Nikfarjam et al, 2015). They always consider different features such as N-grams, POS tags, negation, sentiment word, polarity and etc. These methods could offer a reasonable accuracy, however they are built with supervised training and require large volume of data during the learning process which requires a tremendous amount of manual effort. Various architectures of neural network have also been researched for the detection of ADRs. People have tried convolutional neural network(Lee et al, 2017), recurrent neural network(Cocos et al, 2017) or combine them together(Huynh et al, 2016). Moreover, attention mechanism and CRF are sometimes added into the architecture to improve the performance of system(Pandey et al, 2017).

Although there is substantial previous research on ADRs extraction from English online forums, very limited research was done on Chinese data. To the best of our knowledge, this paper is the first attempt to mine ADRs from two popular Chinese social media sites, namely Xunyiwenyao [2], Haodaifu [3]. Xunyiwenyao and Haodaifu are both online public forums for health-related discussions.

Herein, we propose a semi-supervised learning framework requiring very little manual annotations for mining ADRs from Chinese social media. As an

[2] http://club.xywy.com/
[3] http://www.haodf.com/

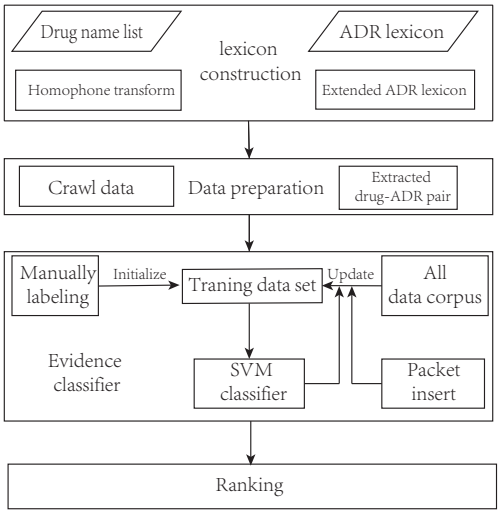

Fig. 1: System framework

alternative to the methods described above, we build a list of commonly misspelled drug names and extend the customized lexicon with colloquial words and adjective modifiers, in order to address the problem of irregular ADR terms and typos. We also focus on distinguishing between indications and ADRs by training a binary classifier, using the SVM model. To train the classifier, we introduce an automatic labeling algorithm to generate large amount of training data.

2 Methods

Our framework (depicted in Fig. 1) is divided into four parts, namely constructing lexica, extracting candidate ADRs, classifying evidences and finally ranking the ADRs.

2.1 Lexicon construction

We need two lexicons, one for the names of medications of interest; the other for ADRs to be recognized from text.

2.1.1 Lexicon of medication

We start with a list that contains common names and registered trade names of known drugs. On social media, drug names may be spelled with variation, either by similar characters or homophones. For example, a drug called "耐信(Nexium)"(nài xìn in Chinese phonetic alphabet) may be misspelled as "奈

Table 1: ADRs lexicon

5'-核苷酸酶下降(5'-nucleotidase decline)	各种肝功能分析(Variety of liver function)	肝胆系统检查(Hepatobiliary system check)	各类检查(Various types of inspection)
5'-核苷酸酶增加(5'-nucleotidase increase)	各种肝功能分析(Variety of liver function)	肝胆系统检查(Hepatobiliary system check)	各类检查(Various types of inspection)
A型肝炎(Hepatitis A)	各种肝脏病毒感染(Various liver virus infection)	肝脏及肝胆类疾病(Liver and hepatobiliary diseases)	肝胆系统疾病(Hepatobiliary system diseases)
BK病毒感染(BK virus infection)	多瘤病毒感染(Polyomavirus infection)	传染性病毒感染(Contagious viral infection)	感染及侵染类疾病(Infection and infection diseases)

信"(nài xìn), "乃信"(nǎi xìn) and so on. To solve this problem, we expand each correct character in a drug name to several commonly misspelled characters in Chinese according to the Chinese phonetic alphabet. For example, "耐" is extended to "奈" or "乃", while "信" is extended to "心", "新" and so on. However, if "耐信" is transformed to "耐心", which is a commonly used Chinese word, many irrelevant posts containing "耐心" maybe returned. Thus common Chinese words which are clearly not drug names are filtered out. After this kind of expansion, we obtain a total of 110779 different drug names for 79 drugs of interest. The list of all these 79 drugs of interest can be found in Appendix A.

2.1.2 Basic ADR lexicon

The basic ADR lexicon comes from four sources: NCI Common Terminology Criteria for Adverse Events (CTCAE) (Trotti et al, 2003), Sougou Pinyin ADRs lexicon[4], MedDRA(The Medical Dictionary for Regulatory Activities) (Brown et al, 1999) and the ADR database by Ye et al (Ye et al, 2014). CTCAE contains formal terms of the ADRs used for adverse event reporting to regulatory agencies. Sougou ADRs is utilized particularly for colloquial terms. Here are some examples: "听力降低"(poor hearing), "焦急不安"(anxious), "健忘"(forgetful), "头发稀疏"(hair thinning). Both CTCAE and Sougou ADRs are available in Chinese. The ADRs database covers more than 6000 ADRs in English. It was translated into Chinese by Google Translate[5]. In addition, classification of these terms is very important. Because some words have the same or similar meaning, their results can be merged in the following analysis

[4] Sogou Pinyin is a Chinese input method, and there are many available lexicons, one of which is the ADRs lexicon: http://pinyin.sogou.com/dict/detail/index/644 .

[5] https://translate.google.com/

steps. For example, “体重减少” (loss of weight) is the same as “体重下降” (drop in weight). If we classify both words in the same category, their result can be directly added and we get one total result for later discussion. Finally, based on MedDRA’s category, we classify all the words into structured lexicon which has four levels. The lowest level contains ADR words from the three data sources. The three upper levels are custom categories in MedDRA. In Table 1, the first column in the left is the fourth level and the next three columns are the upper levels in MedDRA.

2.1.3 Extended ADR lexicon

To improve the ability to match colloquial terms in online discussion, we further expand our basic ADR lexicon by adding variations of the terms. For example, when a person has a headache, he or she may say “头痛(headache)” or “头有点痛(got a little headache)”, the latter of which is a slight variation with a degree modifier between an organ name and symptom word such as “痛” (pain), and is added to our extended lexicon.

There is a variety of such degree modifiers. We adopt a data-driven approach to mine such degree modifiers by pattern-matching an organ name, up to 5 characters and a symptom word, for example “头(head)XXXXX 痛(pain)”, from online discussion corpus. The algorithm to extend ADR lexicon is presented briefly as Algorithm 1.

Algorithm 1 Extending ADR lexicon

```
//Construct regular expression patterns
for each term in basic ADRs do
    if term contains organ then
        construct a regular pattern
//Discover degree words
for each line in all data do
    if line match a pattern then then
        count one for this word
//Extend lexicon
for each term in lexicon do
    if term contains organ then
        for each word in words list do
            insert word into term to generate a new term
```

2.2 Data sources and data preparation

This section describes two Chinese social media and how we extract evidences of ADRs for drugs from them. Besides Haodf and Xywy, we have also tried to use the data from Weibo[6], discussions about it could be seen in 2.2.2.

[6] http://weibo.com

Fig. 2: Question posted on Xunyiwenyao website

2.2.1 Chinese social media

Xunyiwenyao was established in 2004. By 2014, it has over 80,000,000 registered accounts, over 20,000,000 daily independent, and is ranked first in the medical and health service industry. The forum contains 14 categories and 64,050 discussion threads on average, every day. Each discussion thread starts with a patient's question, which is followed by responses from multiple doctors or other patients (see Fig. 2).

Haodaifu was launched in 2006. Its physician-patient interactive forum is the largest in China, with over 501,000 registered healthcare professionals. It contains 29 categories and 18,632,602 discussion threads until now. The format of the discussion is similar to Xunyiwenyao.

2.2.2 Issues with Weibo posts

Weibo is a Chinese microblogging website where a user can start a new conversation in any topic upon which their friends may respond with comments or forward the discussion to other people.It was established in 2009. By 2016, it has over 297,000,000 subscribers and 132,000,000 daily users[7]. The number of posts each day is around 100,000,000[8].

Weibo messages are terse and informal but the quality of such messages is lower than the first two data sources while the quantity is much larger.

The quality of Weibo posts is lower because:

– A doctor often post a message on Weibo after answering a question in Xunyiwenyao or Haodaifu, so some of the data from Weibo is redundant;

[7] http://www.businessofapps.com/sina-weibo-revenue-and-statistics/
[8] www.useit.com.cn/thread-14392-1-1.html

- When users comment and forward a message, it rarely contains a complete sentence, which means it's highly dependent on the original message and makes it harder to processing;
- Very few messages are really about ADRs. For example, there are 7734 messages about Betaloc that we crawled from Weibo, but only 1323 messages contain both Betaloc and a condition;
- There is lots of noise, such as commercial advertisements. In the previous example, out of 1323 messages containing both Betaloc and a condition, only 36% of the messages are really experience reports from the patients who have taken Betaloc.

In consequence, we do not use the data from Weibo. We only use the combination of data from "Xunyiwenyao" and "Haodaifu" to find the potential ADRs for the 79 drugs of interest.

2.2.3 Extraction of evidences

First, we preprocess all the user posts from three websites. If one post contains a drug name of interest, this post is considered as an "effective" target. All sentences in "effective" posts are segmented by ICTCLAS (Zhang et al, 2003), a Chinese word segmentation tool.

With the ADR lexicon, we can detect candidate ADR terms from the effective posts. However, when a drug name X is mentioned in a post, the user may not actually have taken that drug. Similarly, when an ADR term is mentioned, the user may not actually have the symptom, or the symptom may not be the result of taking X. Therefore, given a pair of a drug name and an ADR, we need to determine whether the ADR is truly the consequence of taking the drug, given the context of the pair in the post. Because of that a drug-ADR pair that is too far away from each other in the text is not reliable, the context is defined as one or more consecutive sentences where the distance between drug and ADR is less than 55 Chinese words (including punctuations but excluding spaces). We ensure that each context contain one drug-ADR pair.

We define a context as a positive evidence if the candidate ADR in the context is a real ADR, while the other cases belong to the negative sentence. The following are two contexts showing a positive evidence and a negative evidence:

- 服用易瑞沙后头痛，眼睛复视，模糊 (After taking Iressa, had a headache, eye diplopia and blurred vision)
- 吃的是奥美拉唑，克拉霉素，阿莫西林，吗丁啉等药，咳嗽有所减少 (After taking Omeprazole, Clarithromycin, Amoxicillin, Domperidone and other drugs, cough lessened)

2.2.4 Data set

We have crawled user messages posted between January 2011 to April 2015 on Haodaifu and Xunyiwenyao. These messages mentioned 79 drugs, which

Table 2: Category of drugs studied

Category	Number of drugs	Diseases	Number of drugs
Hypertension	29	Hyperacidity	2
Diabetes	18	Lung cancer	1
Asthma	15	Rhinitis	1
Statins	9	Schizophrenia	1
Breast cancer	1	Acute coronary syndrome	1
Anesthesia	1		

treat 11 types of diseases. Table 2 summarizes the diseases and the number of corresponding drugs. In total, 456,753 posts were crawled.

After preprocessing these posts, we obtain 302,180 sentences where a drug-ADR pair is revealed. We first manually label 1200 sentences which contains 600 positive evidences and 600 negative evidences. Then we divide them into training set, tuning set and test set. Finally, we get a training set with 300 positive evidences and 300 negative evidences, a tuning set with 200 positive evidences and 200 negative evidences and a test set with 100 positive evidences and 100 negative evidences.

2.3 Evidence Classifier

Given a drug name and a medical condition, identified by the extended lexicon, as well as their context in the original text, the problem of evidence classification is to determine whether the medical condition is actually an ADR resulting from the drug. Next we present a method to train such an evidence classifier. In particular, we show how to produce large amount of training data by automatic labeling.

2.3.1 Building the training set

A supervised classifier requires labeled training data. However, manual labeling on user discussion posts can't scale up because of the large amount of informal use of language and colloquial terms. Fortunately, information in the package insert of the drugs, e.g., the indications and the known side effects of the drug, can be used to automatically generate labeled data.

Our first and simple idea is to regard a pair of drug and medical condition as true if the medical condition is listed as a side effect in the package insert of the drug. Conversely, we regard the pair as false if the medical condition is listed as an indication of the drug. All other pairs are discarded from labeled data set. However, this approach is not perfect. For example, "头晕(dizzyness)" is a known ADR for Betaloc, but sometimes in the real discussion it serves as an indication:

– 突然感到头晕心慌,坐卧不安,去医院检查血压160.100 心电图心动过速160次,开了倍他乐克(Suddenly I felt dizzy, flustered, and restless, my blood pressure

was at 160/100; tachycardia electrocardiogram was at 160 times. Consequently I was given Betaloc)

Similarly, "房颤(atrial fibrillation)" is an indication for Betaloc, but sometimes it is reported as if it's a side effect:

– 后根据医嘱，可达龙减至1/4片每天，加服倍他乐克缓释片一片。一段时间后出现房颤(According to the doctor's advice, Cordarone was reduced to 1/4 tablets per day, plus one tablet of Betaloc(slow release). Atrial fibrillation occurred after a period of time)

Because the actual situation arising from patients' experience may be more complicated than specified on the inserts, we adopt a semi-supervised approach instead. We first use the 600 manually labeled data to train a simple SVM classifier and use it to predict for all the sentences in the corpus. The features used are discussed in Section 2.3.2. If the classifier predicts a sentence to be positive, and the medical condition is a known ADR for the drug according to the insert, we add this sentence into the new positive training set. If a sentence is predicted to be negative, and the condition in that sentence is a known indication of the drug, then we add this sentence into the negative training set. We exclude those sentences for which the prediction of classifier and content of the package insert are different. The new training set also contains our original 600 manual labeling data.

With little manual effort, we have now obtained a much larger set of positive and negative training data (called semi-supervised data) — 12,238 training instances in total. By manual validation, the accuracy of such automatic labeling is 82%.

2.3.2 Features extraction

Our main evidence classifier extracts the following features (see Table3), after parsing the evidence sentences into dependency trees:

The set of features described above are used in both the initial and the final classifier. However, with more training data, the final classifier can better distinguish unseen tokens. It's worth noting that all these seven features are independent of the name of the drug and the ADR.

2.3.3 Automatic labeling by bootstrapping

We choose SVM as our primary classifier, because our feature vectors are high-dimensional (many different words). The overall process of our method is indicated in Algorithm.2.

The above algorithm uses the package inserts and the initial classifier M' to generate more training data. One interesting thought is to use that newly obtained classifier M to label even more training data, and thus build a newer classifier. This process can go on iteratively until no more new training data is obtained. We will show the results of this in Section 3. The training data

Table 3: Features that we extracted

Notation	Description	Examples
Feature 1	Verbs before the drugs	"服用(take)" in "服用倍他乐克(take *Betaloc*)"
Feature 2	Verbs before the conditions	"感到(feel)" in "感到头晕(feel dizzy)
Feature 3	Verbs after the conditions	"好转(improved)" in "头疼好转(headache improved)"
Feature 4	Preposition, conjunction and noun of locality	"因为(because of)" in "因为头疼(because of headaches)" and "后(after)" in "服用倍他乐克后(after taking *Betaloc*)"
Feature 5	Punctuations that surround drugs and conditions	"，" and "。" in "吃完后，感到头疼。(feel headache after eating)"
Feature 6	The number of other drugs and other conditions between the drug and condition of interest	Both numbers are equaling to 1 in the sentence "服用信必可和舒利迭之后，感到头痛，身上有些地方还有荨麻疹(After taking *Symbicort* and *Seretide*, feel headache, there also appears urticaria in some places) if the drug and condition of interest is "信必可(Symbicort)" and "荨麻疹(urticaria)"
Feature 7	A boolean value that indicates whether condition appears in front of the drug or not	"true" in "因为哮喘，医生开了信必可(Because of asthma, the docter prescribed *Symbicort*)" and "false" for the sentence "用信必可来治疗哮喘(use *Symbicort* to treat asthma)"

Algorithm 2 Automatic labeling by bootstrapping

```
Manually label small amount of seed data S
Train an initial SVM classifier M from S
Calculate F1-score of this SVM classifier based on the test data set
repeat
    //Use M to classify all the sentences and enlarge our training set with the help of packet inserts
    for each sentence in corpus do
        if M predicts this sentence to be positive && the medication condition is a known ADR for the drug according to the packet insert then
            Add this sentence to the positive training set
        else if M predicts this sentence to be negative && the medication condition is a known indication of the drug according to the packet insert then
            Add this sentence to the negative training set
        else keep this sentence in the corpus
    //update the SVM classifier
    Use the new training set to train a new SVM classifier and update M
    Calculate F1-score of the updated classifier M based on the test data set
until F1-score converge
```

obtained at the final iteration is called semi-supervised data and will be used to train our SVM classifier and the other baseline classifiers (see Section 2.4).

2.4 Baseline classifier techniques

2.4.1 Pattern-based method

Beside the above semi-supervised learning method, we have also tried a intuitive pattern-based classifier as a baseline. We extract preposition, conjunction and noun of locality from sentences as patterns from training data generated by package inserts. Each pattern has a weight, which is its frequency of occurrence; a negative pattern extracted from negative examples will have a negative weight. For example, below are two patterns we extracted and their weight:

- drug ... 后 ... adr ... 20
- adr ... 后 ... drug ... -3

For a new sentence that can be matched to several patterns, the score is the sum of these patterns. Then a classifier is built based on the score: if the score is greater than 0, it's positive; otherwise negative.

2.4.2 HMM-based classifier

We train a HMM classifier (Sampathkumar et al, 2014). Particularly, comparing to original HMM paper where the sentences to be classified may not contain a drug-ADR pair, our task is more challenging because we firstly ensure a drug-ADR pair in all sentences and then make the classification. We train two HMM classifiers in all. One classifier is only trained with 600 manually-labeled data and another classifier is trained with the semi-supervised data by using the package insert.

2.4.3 CRF-based classifier

We train a CRF-based classifier (Nikfarjam et al, 2015). We also use two kind of data to train the two CRF-based classifiers: one with 600 manually-labeled data and another with semi-supervised data.

Both the HMM and CRF classifiers were slightly modified to adapt to the Chinese input. For example we use ICTCLAS to segment and POS to tag the input sentences.

2.5 Ranking

For each drug, there are many candidate ADRs. We are interested in those of high confidence. One way of ranking the ADRs of a drug is by the number of its

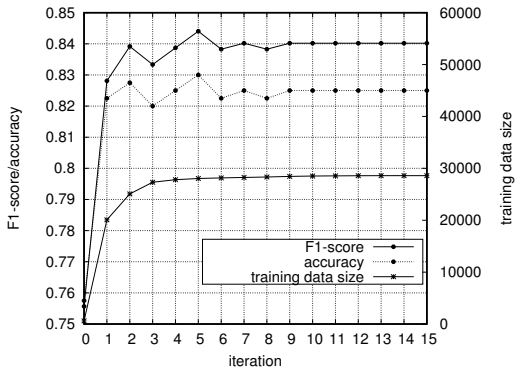

Fig. 3: F1-score, accuracy and training data size of the new SVM classifier at each iteration

appearances in positive evidence posts. This doesn't work well because, most discussions about a drug involves the indications of the drug. For example, discussion about *Betaloc* would naturally include a lot of occurrences of the term "hypertension" and the absolute number of such mentions is very large. Although our classifier can give a high accuracy, a number of sentences which contains "hypertension" as ADR are incorrectly predicted to be positive. Consequently, "hypertension" would be ranked highly as an ADR of *Betaloc*. To solve this problem, we rank the ADRs according to the frequency of the positive evidences minus that of the negative evidences. This approach effectively lowers the rankings of the indications of a drug, but promotes real ADRs.

3 Results

We divide our evaluation into six parts. Firstly, we run the automatically labeling algorithm iteratively and show the change of the performance. Secondly, we will examine the importance of different features in the SVM classifier. Thirdly, we compare the accuracy of our final classifier with other several baseline classifiers (HMM, CRF and pattern-based), the difference caused by the difference training set will also be shown. Fourthly, we evaluate the effect of enlarging the drug and ADR lexica. Finally, we evaluate the accuracy of discovered ADRs with the help of drug package inserts, and show the top-ten discovered ADRs of several drugs, as verification and supplement for the known ADRs in the package inserts.

3.1 Impact of the iteration

Fig. 3 shows the accuracies and F1-scores on the tuning set after each iteration, using the bootstrapping approach in Section 2.3. The result at iteration 0 is obtained using only the manually labeled data. After each iteration, the

Table 4: The effectiveness of classification features

SVM Features	positive pairs	negative pairs	R	P	F1	accuracy
All	184/200	148/200	0.92	0.78	**0.844**	**0.830**
without feature 1	175/200	152/200	0.875	0.785	0.827*	0.818
without feature 2	184/200	147/200	0.92	0.776	0.842	0.828
without feature 3	175/200	**153**/200	0.875	**0.789**	0.829*	0.820
without feature 4	**187**/200	144/200	**0.935**	0.770	**0.844**	0.828
without feature 5	169/200	131/200	0.845	0.710	0.772*	0.750
without feature 6	180/200	141/200	0.900	0.753	0.820*	0.803
without feature 7	173/200	146/200	0.865	0.762	0.810*	0.798

training set will enlarge, however the speed of growth becomes slow in each iteration and drops to 0 at 15^{th} iteration. By using the tuning set which contains 400 manually labeled data (200 positive + 200 negative) to calculate the f1-score and accuracy of our SVM classifier in each iteration, we observe quick convergence: the two values keep constant after 9^{th} iteration.

The biggest improvement of performance comes from the 0^{th} iteration to the 1^{st} iteration since the most knowledge is acquired in the first round of bootstrapping. The gain in accuracy and f1-score saturates after a peak is reached at the 5^{th} iteration. We therefore use the training data obtained at that time to train our final SVM classifier and other baseline classifiers.

3.2 The effectiveness of classification features

To examine the contribution of each feature of our SVM classifier, we use the previous tuning set which contains 400 manually labeled sentences to performed ablation tests on the tuning set. The result is shown in Table 4. Compared with All features set, those significant changes (the difference of F1-score is more than 0.10) are marked with asterisks. Besides, the highest values in each column are highlighted in bold.

We find that each feature does the contribution for the performance of the classifier. Among all the features, feature 1, 3, 5, 6, 7 are the most important ones as F1-score decreases sigificantly without these features.

3.3 Drug-ADR association

According to the previous research, we use the training data obtained at the 5^{th} iteration and all the features to train our SVM classifier. To make the comparison with several baseline classifiers, another 200 manually-labeled test data (100 positive + 100 negative), which are different from the previous tuning set, is chosen to check the performance of the various classifier. The result is shown in Table 5. There are three kinds of training data:

Table 5: Performance of various classifier

Methods	positive pairs	negative pairs	Recall	Precision	F1-score
Manual labels (Pattern-based)	24/100	97/100	0.24	**0.889**	0.378
Manual labels (HMM)	62/100	85/100	0.62	0.805	0.700
Manual labels (CRF)	86/100	75/100	0.86	0.775	0.815
Manual labels (SVM)	68/100	87/100	0.68	0.840	0.751
Auto labels from inserts (Pattern-based)	47/100	77/100	0.47	0.671	0.553
Auto labels from inserts (HMM)	85/100	55/100	0.85	0.654	0.739
Auto labels from inserts (CRF)	98/100	32/100	**0.98**	0.590	0.737
Auto labels from inserts (SVM)	81/100	65/100	0.81	0.698	0.75
Semi-supervised labels (Pattern-based)	76/100	89/100	0.76	0.874	0.813
Semi-supervised labels (HMM)	87/100	54/100	0.87	0.654	0.747
Semi-supervised labels (CRF)	98/100	34/100	**0.98**	0.598	0.742
Semi-supervised labels (SVM)	86/100	79/100	0.86	0.804	**0.831**

- **Manual labels:** use the manually labeled training set with 300 positive instances and 300 negative instances
- **Auto labels from insert:** use the training data that we obtained according to the package insert directly without help of the manually labeled data. If the symptom in the sentence is ADR according to the package insert, it will be added into positive training set. Inversely, if the symptom in the sentence is indication according to the package insert, it will be added into negative training set.
- **Semi-supervised labels:** use the training data that we obtained after the 5^{th} iteration.

The pattern-based classifier depends a lot on the size of the training data set. More training data could help it to recognize more patterns of a positive sentence. In consequence, the performance improves a lot when using semi-supervised labels.

The HMM-based classifier emphasizes on the structure of sentences. The performance improved if the structure in training set and testing set is standard. Therefore, when we use the manually-labeled data to train the HMM classifier, the small size of training data set results in a low precision. It can be also seen that the percentage of true positives is inversely correlated with the percentage of true negatives. This means a classifier is biased to produce either more positive labels or more negative labels. A good classifier, such as the one trained with the semi-supervised labels manages to strike a balance between the two biases and produce a better overall F1-score.

Table 6: Enlarging data set through homophone transform

	倍他乐克 (Betaloc)	耐信 (Nexium)	拜唐苹 (Glucobay)	氨茶碱 (Aminophylline)	All 79 drugs
official name	24073	6521	530	7493	158695
homophone	13177	6369	1611	2388	143485
total	37250	12890	2141	9881	302180
%increase	35.4%	49.4%	75.2%	24.2%	47.5%

CRF-based classifier use the sequence labeling with word embedding cluster features, which reduce the effect of the training set's size. However, this kind of classifier also depends on the grammatical form of a sentence. When training set enlarges, the structure of negative instances becomes various and do not have a regular form, which leads to a bad performance of the CRF classifier.

In short, both the HMM and CRF concentrate more on the information of the single word itself and its limited surrounding words. However, SVM focus on the features of the whole sentence.

The semi-supervised data, which is doubly verified by the primary SVM classifier and package inserts, may not have a very standard form (e.g., some sentences do not have the causal keyword but have a lot of noisy words between the ADR and its associated drug). For those user posts, which do not have a standard form, SVM performs clearly better because of its global view, and HMM doesn't perform as well because it requires sentences in their standard form.

3.4 Homophone transformation and extended ADR lexicon

As shown in Table 6, our data set, measured by the number of sentences containing at least one of the 4 selected drugs and an ADR, is enlarged significantly after homophone transformation.

Among all the 302,180 sentences which contains a (drug, ADR) pair, there are totally 1,328 sentences where the candidate ADR contains an adverb of degree and can only be extracted by using the extended ADR lexicon. Although 1,328 is not large compared to 302,180, extended ADR lexicon could also help us to enlarge the data set to find more potential ADRs.

In addition, we randomly select 100 original post to measure the performance of our ADR lexicon. Among all the 451 medications mentioned, we could detect 159 medications. After calculation, we obtain the precision and recall of our ADR lexicon is 1.0 and 0.353. Although there are still a number of undetected colloquial medications, we have tried our best to combine lexicons from sources(see 2.1.2) and add the colloquial term(see 2.1.3).

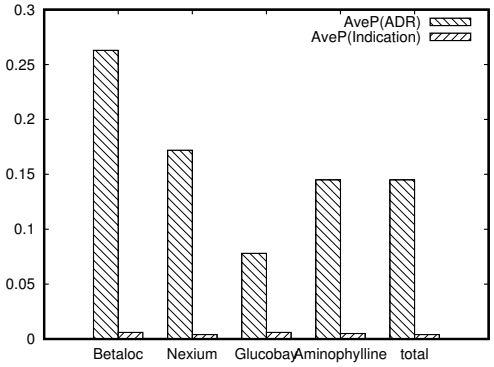

Fig. 4: End-to-end rankings' AveP

3.5 End-to-end ranking

By using the ranking method which is referred in Section 2.5, our system returns a ranked list of possible ADRs when given a drug. We evaluate the end-to-end performance of the system by the Average Precision ($AveP$) according to the package insert of the drug:

$$AveP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{number\ of\ ADRs\ in\ package\ inserts} \tag{1}$$

where $P(k)$ is the precision at cut-off k in the list, $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, 0 otherwise.[9]

We expect the true ADR of a drug to rank high in the list while the true indication ranks lower in the list. The ground truth we use here is the known ADRs and known indications of four random-sampled drugs according to the package inserts. Figure 4 shows the results of the four previous randomly chosen drugs, 倍他乐克($Betaloc$), 耐信($Nexium$), 拜唐苹($Glucobay$) and 氨茶碱($Aminophylline$). We also calculate the weighted average of $AveP$ for all the 79 drugs.

From Fig. 4, we can see that $AveP(ADR)$ is much larger than $AveP(Indication)$, which means that most of ADRs that our classifier discovers are already included in the package insert. Besides, the known indications are not in our returned ADR list or ranked very low in our list.

Together with Table 6, which gives the sizes of the datasets for four drugs, we learn that more data helps to increase the ADR prediction accuracy.

Table 7: Top 10 discovered ADRs for 4 common drugs

药物 (Drugs)	倍他乐克 (Betaloc)	耐信 (Nexium)	拜唐苹 (Glucobay)	氨茶碱 (Aminophylline)
副作用 (ADRs)	咳 嗽(2.45%) (Cough)	咳 嗽(1.77%) (Cough)	不 适(3.31%) (Discomfort)	咳 嗽(51.39%) (Cough)
	紧 张(2.06%) (Nervous)	头 晕(1.09%) (Dizziness)	无 力(2.18%) (Acratia)	头 晕(0.69%) (Dizziness)
	不 适(4.04%) (Discomfort)	不 适(2.30%) (Discomfort)	发 热(1.48%) (Fever)	恶 心(0.57%) (Nausea)
	心 悸(2.82%) (Palpitation)	紧 张(0.32%) (Nervous)	头 晕(2.70%) (Dizziness)	心 悸(0.26%) (Palpitation)
	头 晕(5.52%) (Dizziness)	便 秘(0.85%) (Constipation)	乏 力(1.31%) (Weak)	呕 吐(1.13%) (Emesis)
	疲 劳(0.67%) (Fatigue)	疲 劳(0.16%) (Fatigue)	瘙 痒(0.87%) (Itching)	心 动 过速(0.19%) (Tachycardia)
	头 痛(1.32%) (Headache)	失 眠(0.50%) (Insomnia)	腹 泻(1.13%) (Diarrhea)	心 律 失常(0.26%) (Arrhythmia)
	恶 心(0.89%) (Nausea)	头 痛(0.36%) (Headache)	低 血 糖(3.14%) (Hypo-glycemia)	打 鼾(0.22%) (Snore)
	便 秘(0.16%) (Constipation)	心 悸(0.11%) (Palpitation)	虚 弱(0.52%) (Asthenia)	抽 搐(0.22%) (Tic)
	瘙 痒(0.14%) (Itching)	皮 肤 过敏(0.12%) (Skin allergy)	咳 嗽(0.61%) (Cough)	紧 张(0.12%) (Nervous)

3.6 Top-ten discovered ADRs

Table 7 shows the top-ten discovered ADRs for 4 aforementioned drugs. The number in the parentheses the percentage which is calculated as followed:

$$percentage = \frac{\#\ of\ patients\ who\ report\ that\ ADR}{\#\ of\ posts\ which\ discuss\ this\ drug} \tag{2}$$

ADRs which don't have direct match in the package inserts (therefore potentially new discoveries) are marked in *red*.

In Table 7, we discovered many ADRs that are already included in the package inserts. Although these ADRs are known, the frequency statistics can be valuable for: i) verifying ADRs listed in the package inserts; ii) studying the relative frequency between the ADRs. For example, the frequency of *Fatigue* and *Constipation* of *Betaloc* in package insert are both larger than 1%, but they are 0.67% and 0.16% respectively in our result.

There are also a number of ADRs without direct match in the manuals. These fall into several cases:

[9] *AveP* is defined at https://en.wikipedia.org/wiki/Information_retrieval

Newly discovered ADRs (e.g., "咳嗽(Cough)" for "倍他乐克(*Betaloc*)"). This is the most valuable discovery for the drug maker in the analysis of the drug reactions in perhaps a small population previously not considered.

Synonyms of the known ADRs (e.g., "疲乏(Exhaustion)" is a synonym of "疲劳(Fatigue)" for "耐信(*Nexium*) ". While they are synonyms, the ADRs listed in package inserts are often some terminologies and the colloquial synonyms can help patients understand them easily.

Generalization of the known ADRs (e.g., "呕吐(Emesis)" is a specialization of the symptom "不适 (Discomfort)" for "倍他乐克 (*Betaloc*)"). Some ADRs from package inserts is a specific symptom. Our results give a general term.

4 Conclusion

We have proposed an effective framework for extracting and analyzing ADRs from Chinese online social media. It uses a lexicon-based method to extract ADRs from the data followed by a binary classifier to identify the positive evidences. In this framework, we introduce a data-driven algorithm to extend the drug and ADR lexica. In order to build the evidence classifier, we propose an automatic labeling algorithm to produce large amounts of labeled sentences. Completely relying on the information from the package inserts produces training data that is too noisy. Our tradeoff is a semi-supervised approach where we manually label a small set, then use these data and package inserts collectively to generate more training data. This approach was shown to be highly effective.

Acknowledgements This work has been supported by AstraZeneca.

References

Benton A, Ungar LH, Hill S, Hennessy S, Mao J, Chung A, Leonard CE, Holmes JH (2011) Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. Journal of Biomedical Informatics 44(6):989–996

Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, Day R, Ferraz MB, Hawkey CJ, Hochberg MC, et al (2000) Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. New England Journal of Medicine 343(21):1520–1528

Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, Lines C, Riddell R, Morton D, Lanas A, et al (2005) Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. New England Journal of Medicine 352(11):1092–1102

Brown E, Wood L, Wood S (1999) The medical dictionary for regulatory activities (meddra). Drug Safety 20(2):109–117

Cocos A, Fiks AG, Masino AJ (2017) Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. Journal of the American Medical Informatics Association p ocw180

Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, Dasgupta N (2014) Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug safety 37(5):343–350

Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, Shoor S, Ray WA (2005) Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. The Lancet 365(9458):475–481

Gurulingappa H, Toldo L, Rajput AM, Kors JA, Taweel A, Tayrouz Y (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

Hahn U, Cohen KB, Garten Y, Shah NH (2012) Mining the pharmacogenomics literature—a survey of the state of the art. Briefings in bioinformatics 13(4):460–494

Harpaz R, Haerian K, Chase HS, Friedman C (2010) Statistical mining of potential drug interaction adverse effects in fda's spontaneous reporting system. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2010, p 281

Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C (2012) Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics 91(6):1010–1021

Huynh T, He Y, Willis A, Rüger S (2016) Adverse drug reaction classification with deep neural networks. COLING

Jiang L, Yang CC, Li J (2013) Discovering consumer health expressions from consumer-contributed content. In: SBP, Springer, pp 164–174

Jonnagaddala J, Jue TR, Dai H (2016) Binary classification of twitter posts for adverse drug reactions. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Big Island, HI, USA, pp 4–8

Karimi S, Kim S, Cavedon L (2011) Drug side-effects: What do patient forums reveal. In: The second international workshop on Web science and information exchange in the medical Web, ACM, pp 10–11

Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G (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, Association for Computational Linguistics, pp 117–125

Lee K, Qadir A, Hasan SA, Datla V, Prakash A, Liu J, Farri O (2017) Adverse drug event detection in tweets with semi-supervised convolutional neural

networks. In: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp 705–714

Li YA (2011) Medical data mining: Improving information accessibility using online patient drug reviews. PhD thesis, Massachusetts Institute of Technology

Liu X, Chen H (2013) Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In: International Conference on Smart Health, Springer, pp 134–150

Liu X, Liu J, Chen H (2014) Identifying adverse drug events from health social media: a case study on heart disease discussion forums. In: International Conference on Smart Health, Springer, pp 25–36

Nikfarjam A, Gonzalez GH (2011) Pattern mining for extraction of mentions of adverse drug reactions from user comments. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2011, p 1019

Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G (2015) Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association 22(3):671–681

Pandey C, Ibrahim Z, Wu H, Iqbal E, Dobson R (2017) Improving rnn with attention and embedding for adverse drug reactions. In: Proceedings of the 2017 International Conference on Digital Health, ACM, pp 67–71

Sampathkumar H, Chen Xw, Luo B (2014) Mining adverse drug reactions from online healthcare forums using hidden markov model. BMC medical informatics and decision making 14(1):91

Sarker A, Gonzalez G (2015) Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics 53:196–207

Sharif H, Zaffar F, Abbasi A, Zimbra D (2014) Detecting adverse drug reactions using a sentiment classification framework. In: SocialCom, Academy of Science and Engineering (ASE), USA,© ASE 2014

Sohn S, Kocher JPA, Chute CG, Savova GK (2011) Drug side effect extraction from clinical narratives of psychiatry and psychology patients. Journal of the American Medical Informatics Association 18(Supplement_1):i144–i149

Trotti A, Colevas AD, Setser A, Rusch V, Jaques D, Budach V, Langer C, Murphy B, Cumberlin R, Coleman CN, et al (2003) Ctcae v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment. In: Seminars in radiation oncology, Elsevier, vol 13, pp 176–181

Wang W, Haerian K, Salmasian H, Harpaz R, Chase H, Friedman C (2011) A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from pubmed citations. In: AMIA annual symposium proceedings, American Medical Informatics Association, vol 2011, p 1464

Warrer P, Hansen EH, Juhl-Jensen L, Aagaard L (2012) Using text-mining techniques in electronic patient records to identify adrs from medicine use. British journal of clinical pharmacology 73(5):674–684

Wu H, Fang H, Stanhope SJ (2012) An early warning system for unrecognized drug side effects discovery. In: Proceedings of the 21st International Conference on World Wide Web, ACM, pp 437–440

Wu H, Fang H, Stanhope S, et al (2013) Exploiting online discussions to discover unrecognized drug side effects. Methods Inf Med 52(2):152–9

Yang C, Srinivasan P, Polgreen PM (2012a) Automatic adverse drug events detection using letters to the editor. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2012, p 1030

Yang CC, Jiang L, Yang H, Tang X (2012b) Detecting signals of adverse drug reactions from health consumer contributed content in social media. In: Proceedings of ACM SIGKDD workshop on health informatics

Yates A, Goharian N (2013) ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. Springer Berlin Heidelberg

Ye H, Liu Q, Wei J (2014) Construction of drug network based on side effects and its application for drug repositioning. PloS one 9(2):e87,864

Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R (2014) A pipeline to extract drug-adverse event pairs from multiple data sources. BMC medical informatics and decision making 14(1):13

Zhang HP, Yu HK, Xiong DY, Liu Q (2003) Hhmm-based chinese lexical analyzer ictclas. In: Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17, Association for Computational Linguistics, pp 184–187

A List of 79 Drugs Studied

Category	Drug Name	English Name	Manufactor	Total Num of posts
鼻炎 (Rhinitis)	雷诺考特	Rhinocort	AstraZeneca	8164
肺癌 (Lung Cancer)	易瑞沙	Iressa	AstraZeneca	16481
	倍他乐克	Betaloc	AstraZeneca	37250
	波依定	Plendil	AstraZeneca	7089
	颉沙坦	Valsartan	Norvatis	2468
	乌拉地尔	urapidil	Nycomed GmbH	151
	替米沙坦	Telmisartan	Boehringer Ingelheim	1949
	瑞泰	Tritace	Sanofi-Aventis	380
	雅施达	Acertil	LES LABORATOIRES SERVIER	1133
	科素亚	Cozaar	Merck Sharp & Dohme Limited	2853
	海捷亚	Hyzaar	Merck Sharp & Dohme Limited	613
高血压 (Hypertension)	赖诺普利	lisinopril	AstraZeneca UK Limited	287

	再宁平	Zanidip	Recordati S.P.A.	75
	乐息平	Lacipil	GLAXOSMITHKLINE	693
	马来酸伊索拉定	Gaslon N	Nippon Shinyaku Co.,Ltd.	29
	安博维	APROVEL	Sanofi Pharma Bristol-Myers Squibb SNC	2522
	寿比山	Indapamide	Servier	1773
	达爽	Tanatril	天津田边制药有限公司	386
	蒙诺	Monopril	中美上海施贵宝制药有限公司	1222
	多沙唑嗪	Cardura XL	Pfizer Pharma GmbH	229
	合心爽	Altiazem	天津田边制药有限公司	1522
	卡维地洛片	Carvedilol	ROCHE S.P.A.	562
	必洛斯	Blopress	Takeda Pharmaceutical Company Limited	523
	康忻	Concor	Merck Serono GmbH	3104
	贝尼地平	Coniel	Kyowa Hakko Kirin Co.,Ltd.	180
	阿替洛尔	Atenolol	AMRI India Pvt. Ltd.	877
	尼群地平	Nitrendipine	Alvogen Malta Operations Ltd	874
	阿尔马尔	Almarl	Dainippon Sumitomo Pharma Co., Ltd.	901
	络活喜	Norvasc	Pfizer Australia Pty Limited	4636
	锐思力	Rasilez	Novartis Pharma Schweiz AG	2
	特拉唑嗪	Terazosin	Abbott	1316
他汀类药物 (Statins)	可定	Crestor	AstraZeneca	2179
	阿伐他汀	Lipitor	Pfizer Ireland Pharmaceuticals	134
	辛伐他汀	Simvastatin Tablets	Merck Sharp & Dohme (Australia) Pty. Ltd.	1140
	普伐他汀	Pravastatin	华北制药股份有限公司	110
	洛伐他汀	Lovastation	AstraZeneca	751
	氟伐他汀	Fluvastatin	Novartis	267
	葆至能	VYTORIN	MSP Singapore Company,LLC	7
	匹伐他汀	LIVALO KOWA	Kowa Company, Ltd.	57
	氨氯地平阿托伐他汀	Amlodipine Besylate and Atorvastatin Calcium Tablets	Pfizer Inc.	85
胃酸过多 (GERD)	洛赛克	Losec	AstraZeneca	41,957
	耐信	Nexium	AstraZeneca	12,890

急性冠脉综合征 (Acute coronary)	倍林达	BRILINTA	AstraZeneca	179
精神分裂 (Schizophrenia)	思瑞康	Seroquel	AstraZeneca	10,859
麻醉 (Sedation)	得普利麻	Diprivan	AstraZeneca	578
乳腺癌 (Breast Cancer)	瑞宁得	ARIMIDEX	AstraZeneca	1915
糖尿病 (Diabetes)	安立泽	Onglyza	Bristol-Myers Squibb Company	269
	百泌达	BYETTA	Eli Lilly Nederland B.V.	198
	伏格列波糖	Voglibose	Ranbaxy Laboratories Limited	419
	维格列汀	Galvus	Novartis Europharm Ltd.	114
	捷诺维	JANUVIA	Merck Sharp & Dohme (Australia) Pty Ltd	208
	罗格列酮	Avandamet	GlaxoSmithKline	449
	瑞格列奈片	NovoNorm	Novo Nordisk A/S	1157
	吡格列酮	Actos	Takeda Pharmaceutical Company Limited	822
	赛尼可	Xenical	Roche Pharma(Schweiz) Ltd	993
	那格列奈片	Nateglinide Tablet	北京诺华制药有限公司	273
	诺和力	Victoza	Novo Nordisk A/S	59
	长秀霖	Basalin	甘李药业股份有限公司	530
	来得时	LANTUS	Sanofi-Aventis Deutschland GmbH	1719
	诺和锐	NovoRapid FlexPen	Novo Nordisk A/S	1337
	格列吡嗪控释片	Glucotrol XL	Pfizer Inc.	224
	格列美脲片	Amaryl	Sanofi-Aventis Deutschland GmbH	771
	达美康	Diamicron MR	Les Laboratoires Servier	1675
	拜唐苹	Glucobay	Bayer Vital GmbH	2141
哮喘 (Asthma)	普米克	Pulmicort	AstraZeneca	10621
	信必可	Symbicort	AstraZeneca	8349
	安可来	ACCOLATE	AstraZeneca UK Limited	14
	氨茶碱	Aminophylline	Sannova Co	9881
	沙丁胺醇	Salbutamol	EugenPharm Inc, USA	4028
	美普清	Meptin	中国大冢制药有限公司	4252
	吡嘧司特钾	Pemirolast Potassium Tablets	河北医科大学制药厂	216

	盐酸奥洛他定	Allelock	Kyowa Hakko Kirin Co.,Ltd.	1414
	顺尔宁	Singulair	Merck Sharp & Dohme Australia Pty Ltd	54,621
	阿乐迈	Alomide	s.a. ALCON-COUVREUR n.v.	108
	奥克斯都保	Oxis Turbuhaler	AstraZeneca AB	609
	舒利迭	Seretide	Glaxo Wellcome UK Limited	19147
	依匹斯汀	Alesion	Nippon Boehringer Ingelheim Co.,Ltd.	296
	阿米迪	Amiaid	Nitto Denko Corporation	1601
	帮备	Bambec	AstraZeneca	313
Total				**302,180**