Data and text mining

Automatic Discovery of Adverse Drug Reactions through Chinese Social Media

Quanyang Liu[1*], Meizhuo Zhang[2*], Chen Ge[1], Jiemin Wang[2], Jia Wei[2**], Kenny Q. Zhu[1**]

[1]Dept. CSE, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
[2]R&D Information, AstraZeneca, 199 Liangjing Road, Pudong, Shanghai, 201203, China
*The authors contributed equally to this work. **Corresponding authors

Abstract

Motivation: 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.

Results: 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. With this approach, we discovered a large number of side effects for a variety of popular medicines in real world scenarios.

Availability: A web demo is available at http://adapt.seiee.sjtu.edu.cn/~qyliu/demo/ , which contains the ADRs mined for 46 popular medicines.

Contact: Jia Wei (Jenny.Wei@astrazeneca.com) and Kenny Q. Zhu (kzhu@cs.sjtu.edu.cn)

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 R et al., 2010; Harpaz R et al., 2012; Hahn U et al., 2012; Gurulingappa H et al., 2013). These reports are relatively easy to process due to their strict conformance to the adverse event reporting standards. However, the quantity of such reports is limited. Hence, they 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., 2012) 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 targeted 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 et al., 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 machine learning approaches (Sarker et al., 2015; Lardon et al., 2015). Most prior studies (Leaman et al. 2010; Yang et al. 2012; Benton

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

et al. 2011; Wu et al. 2013; Ytes 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 et al 2013), rule (pattern) based methods (Nikfarjam et al. 2011; Benton et al. 2011; Karimi et al. 2011; Yang et al. 2012); or NLP techniques (Sharif et al. 2014; Sarker and Gonzalez 2015). Moreover, a large number of studies have explored machine learning methods for the extraction of ADRs (see Lardon et al. (2015) and Sharker et al. (2015) for a more comprehensive review). These approaches utilize well-studied machine learning methods, and can offer reasonable accuracy. However, they all require large volume of training data during the learning process, a tremendous amount of manual effort.

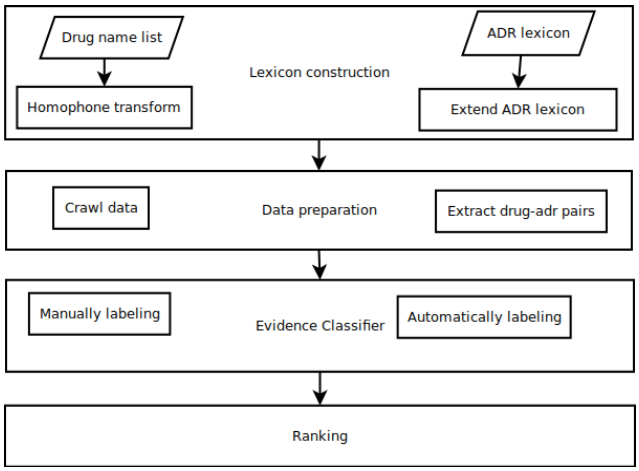

Figure 1 System framework

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 three popular social media sites, namely Xunyiwenyao[2], Haodaifu[3] and Sina Weibo[4]. Xunyiwenyao and Haodaifu are both online public forums for health related discussions. 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.

Herein, we propose a semi-supervised learning framework requiring very little manual annotations for mining ADRs from Chinese social media. As an 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 SVM model. To train the classifier, we introduce an automatic labeling algorithm to generate large amount of training data.

2 Materials and Methods

Our framework (depicted in Figure 1) is divided into four parts, namely constructing lexicons, 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.

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)” may be misspelled as “奈信”, “耐心(patience)”, “乃信” and so on. Coverage is low if only the official drug names are used to search for relevant posts. To solve this problem, we expand each correct character in a drug name to several commonly misspelled characters in Chinese. 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 expansion, we obtain a total of 92858 different drug names for 46 drugs of interest.

Basic ADR lexicon

The basic ADR lexicon comes from four sources: The NCI Common Terminology Criteria for Adverse Events (CTCAE) (Trotti et al., 2003), Sougou Pinyin ADRs lexicon[5] , 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. 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[6]. 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 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. The first column in the left is the fourth level and the next three columns are the upper levels in MedDRA.

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

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

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

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

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 follows.

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)

Algorithm: extend 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
        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 three Chinese social media and how we extract evidences of ADRs for drugs from them. We discuss Weibo separately because the nature of posts on Weibo is substantially different from Xunyiwenyao and Haodaifu.

Chinese social media

Xunyiwenyao was established in 2004. In 2014, there are over 80,000,000 registered accounts, over 20,000,000 daily independent, which earned it the number one ranking in the medical and health service industry [3]. The discussion 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 Figure 2).

Haodaifu was launched in 2006 [4]. Its physician-patient interactive forum is the largest in China, with over 70,000 registered healthcare professionals. It contains 29 categories and 18,632,602 discussion threads until now. The format of the discussion is similar to Xunyiwenyao.

Weibo was established in 2009. By 2015, it has over 222,000,000 subscribers and 100,000,000 daily users[7]. The number of posts each day is around 100, 000,000 [8]. Weibo messages are terse and informal. The quality of such messages is lower than the first two data sources while the quantity is much larger.

Figure 2 Question posted on Xunyiwenyao website

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. So 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. The context is defined as one or more consecutive sentences of up to 50 Chinese words (including punctuations but excluding spaces) that contain a drug-ADR pair. The window size of 50 is commonly used in the literature and should be sufficient to cover 2-3 sentences. A drug-ADR pair that is too far away from each other in the text is not reliable. 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)

We will discuss how to classify evidences into positive and negative ones in the next section.

Issues with Weibo posts

We have mentioned that the discussion volume of Weibo is higher than the other two, but the quality is poorer because:

[7] http://www.businessofapps.com/sina-weibo-revenue-and-statistics/

[8] http://www.bloomberg.com/news/articles/2012-02-28/sina-s-weibo-outlook-buoys-internet-stock-gains-in-n-y-china-overnight

- A doctor would post a message on Weibo after answering a question in Xunyiwenyao or Haodaifu, and which is already contained in the crawled data so it's redundant;
- 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.

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.

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 manually label 400 sentences, from which we extract 211 positive pairs and 211 negative pairs. We then train a simple SVM classifier using this small training set and use the classifier to predict all the sentences in the corpus. The features used are discussed in *Features extraction* section below. If the classifier predicts a sentence to be positive, and the medical condition is a known ADR for the drug according to the manual, we mark this sentence as a positive training instance. If a sentence is predicted to be negative, and the condition in that sentence is a known indication of the drug, then we mark this sentence as a negative training instance.

With little manual effort, we have now obtained a much larger set of positive and negative training data --- 17,382 training instances in total. By manual validation, the accuracy of such automatic labeling is 92%.

Features extraction

Our main evidence classifier extracts the following features, after parsing the evidence sentences into dependency trees (Chang et al., 2009):

- Verbs before the drugs, e.g. "服用(take)" in "服用倍他乐克(take Betaloc)";
- Verbs before the conditions, e.g. "感到(feel)" in "感到头晕(feel dizzy)";
- Verbs after the conditions, e.g. "好转(improved)" in "头疼好转(headache improved)";
- Preposition, conjunction and noun of locality, e.g. "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)";
- Punctuations that surround drugs and conditions;
- The number of other drugs and other conditions between the drug and condition of interest;
- A Boolean value that indicates whether condition appears in front of the drug or not.

The verbs are hard to extract without parsing the sentences. They often occur along with modifiers in the Chinese language. For example, "头疼好转(headaches improved)" would often be expressed as "头疼稍微好转(headaches improved a little bit)", and with the dependency tree we can extract "好转(improved)" from it easily. 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.

Overall flow

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

1. Manually label small amount of seed data S;
2. Train an initial classifier M' from S;
3. Use M' and package inserts to generate more training data T;
4. Train the final classifier M from T.

Bootstrapping of automatically labeling

The above method uses the package inserts and an SVM classifier to generate more training data. One interesting thought is to use that newly obtained classifier 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.

Alternative pattern based method (baseline)

Beside the above semi-supervised learning method, we have also tried a naïve 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 ... 4
- adr ... 后 ... drug ... -4

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 Ranking

For each drug, there are many candidate ADRs. We are interested in ADRs of high confidence. One way of ranking the ADRs of a drug is by the number of its 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". The absolute number of such mentions is very large, and 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 three parts. First, we evaluate the accuracy of the classifier, by showing the accuracy of prediction of drug-ADR association. Then we run the automatically labeling algorithm iteratively and show the change of the F1-scores. Finally, we use MRRs (Mean Reciprocal Rank) of ADRs and indications of package inserts to evaluate the end-to-end results. After evaluation of the platform, we show the top-ten discovered ADRs of several drugs, as verification and supplement for the known ADRs in the package inserts.

Table 2 List of diseases and number of drugs studied

Diseases	Number of drugs	Diseases	Number of drugs
Hypertension	16	Hyperacidity	1
Diabetes	10	Lung cancer	1
Asthma	9	Stomach disease	1
Statins	3	Rhinitis	1
Breast cancer	1	Schizophrenia	1
Anesthesia	1	Acute coronary syndrome	1

3.1 Data set

We have crawled user messages posted between January 2011 to April 2015 on Haodaifu and Xunyiwenyao. These messages mentioned 46 drugs, which treat 12 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 170,196 drug-ADR pairs. We manually labeled 211 positive pairs and 211 negative pairs from 200 random sentences as our seed training data.

3.2 Drug-ADR association

For the test data set, we manually label 200 pairs of drug and ADR, i.e.,100 positive pairs and 100 negative ones. Then we compare the classifier trained from three different labeled data: i) the 422 manually labeled pairs, ii) the labels generated directly from the package insert, and iii) semi-supervised labels. In addition to the SVM classifier, we experimented with three other baseline approaches, namely an HMM (Hidden Markov Model) based classifier (Sampathkumar et al., 2014), a CRFs (Conditional Random Fields) based classifier (Nikfarjam et al., 2015) and the pattern based classifier (see Section 2.3). 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. The result is shown in Table 3.

Table 3 Accuracy of prediction of drug-condition associations

	Positive pairs	Negative pairs	F1
Manual labels	43/100	89/100	0.558
Auto labels from inserts	57/100	83/100	0.655
Semi-supervised labels	**82/100**	**85/100**	**0.837**
HMM	24/100	83/100	0.340
CRFs	37/100	90/100	0.503
Pattern-based	26/74	93/100	0.364

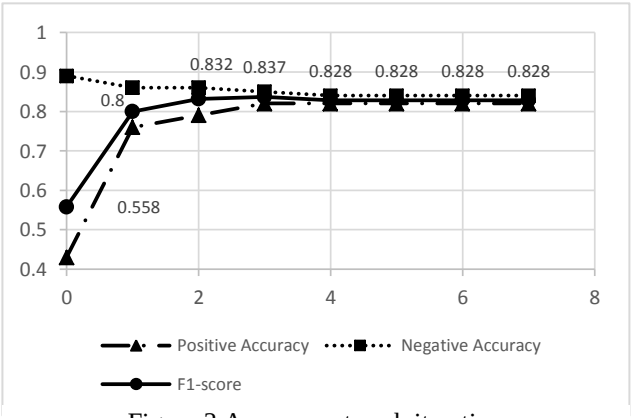

Figure 3 Accuracy at each iteration

Figure 3 shows the accuracies and F1-scores after each iteration using the training data bootstrapping approach in Section 2.3. We observe quick convergence: no extra labeled data is generated after 7th iteration. There is a dramatic improvement in accuracy from the 0th iteration to the lst since the most knowledge is acquired in the first round of bootstrapping. The gain in accuracy saturates after a peak is reached at the 3rd iteration. We therefore use the training data obtained at that time to train our final classifier.

3.3 End-to-end ranking

Table 4 End-to-end rankings' MRR

	易瑞沙(Iressa)		耐信(Nexium)		波依定(Plendil)	
	ADRs	Indications	ADRs	Indications	ADRs	Indications
Manually label	0.021	0.003	0.014	**0.002**	0.055	**0.003**
Label only with package inserts	**0.035**	0.003	**0.022**	**0.002**	0.046	**0.003**
Semi-supervised labels	0.027	0.003	0.015	**0.002**	**0.072**	**0.003**
Patterns method	0.024	0.003	0.009	0.003	0.025	0.004

Because our system returns a ranked list of possible ADRs given a drug, we evaluate the end-to-end performance of the system by the mean rank reciprocal (MRR):

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q.[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 is the known ADRs and known indications of three well known drugs, namely, 易瑞沙(Iressa), 耐信(Nexium) and 波依定(Plendil). To do this evaluation, our classifier was trained using the data for all other drugs. Table 4 shows the results.

3.4 Top-ten discovered ADRs

We show the most frequently reported ADRs (with percentages of their occurrences in the related posts in parenthesis) for 4 most discussed drugs for different indications in Table 5. ADRs that don't have direct match in the package inserts are marked in red.

Table 5 Top 10 discovered ADRs for 4 common drugs

药物 (Drugs)	耐信 (Nexium)	倍他乐克 (Betaloc)	易瑞沙 (Iressa)	思瑞康 (Seroquel)
副作用(ADRs)	头晕(0.55%) (Dizziness)	恶心(0.65%) (Nausea)	皮疹(1.17%) (Rash)	嗜睡(1.78%) (Drowsiness)
	抑郁(0.09%) （Depression）	耳鸣(0.25%) (Tinnitus)	腹泻(0.95%) (Diarrhea)	头晕(1.39%) (Dizziness)
	失眠(0.27%) (Insomnia)	疲劳(0.33%) （Fatigue）	恶心(0.68%) (Nausea)	口干(0.48%) (Dry mouth)
	口干(0.21%) (Dry mouth)	眩晕(0.22%) (Dizziness)	呕吐(0.93%) (Vomit)	恶心(0.51%) (Nausea)
	皮肤过敏 (0.05%) (Skin allergies)	腹痛(0.08%) (Stomach ache)	头晕(0.54%) (Dizziness)	便秘(0.76%) (Constipation)
	眩晕(0.06%) (Dizziness)	嗜睡(0.11%) (Drowsiness)	瘙痒(0.40%) (Itching)	呕吐(0.37%) (Vomit)
	药物过敏 (0.06%) (Drug allergy)	视力模糊 (0.05%) (Blurred vision)	乏力(0.48%) (Weakness)	疲倦(0.15%) (Tired)
	咽喉痛(0.03%) (Sore throat)	瘙痒(0.08%) (Itching)	口腔溃疡 (0.25%) (Mouth ulcers)	呼吸困难 (0.12%) (Difficulty breathing)
	全身乏力 (0.05%) (Malaise)	便秘(0.09%) (Constipation)	头痛(0.46%) (Headache)	耳鸣(0.19%) (Tinnitus)
	鼻塞(0.03%) (Stuffy nose)	黑便(0.03%) (Melena)	厌食(0.22%) (Anoresia)	贫血(0.08%) (Anemia)

[9] https://en.wikipedia.org/wiki/Mean_reciprocal_rank

4 Discussion

As shown in Table 3, The semi-supervised labeling approach provides the best results with F1-score significantly higher than the other approaches. The HMM based method performs the worst, because it only utilizes the positive training data. As a result, the training data is only half of what's used by the other methods. 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.

In order to show the efficiency of end-to-end results of our approach, we calculate the MRRs for ADRs and indications, as shown in Table 4. We can see that our semi-supervised labeling method outperforms both the manually labeling method and the patterns method. However, sometimes it can be worse than labeling only with package inserts. The reason is that only using package inserts the trained classifier is overfitted to the package inserts and it may outperform our semi-supervised labeling method. However, it underperforms when it comes to drug-ADRs association's accuracy.

In Table 5, 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 Tinnitus of Betaloc in package insert is less than 0.1%, but in our result it's 0.25%; the frequency of Stomachache and Constipation of Betaloc in package insert are both larger than 1%, but in our result they are 0.08% and 0.09% respectively.

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

Synonyms of the known ADRs (e.g., "视力模糊(Blurred vision)" is a synonym of "视力损害(Visual impairment)" for "倍他乐克(Betaloc) ". While they are synonyms, the ADRs listed in package inserts are often some terminologies and the colloquial synonyms can help patients understand them easily.

Specialization of the known ADRs (e.g., "失眠(Insomnia)" is a specialized case of "睡眠障碍(Sleep disorders)" for "耐信(Nexium)"). Some ADRs from package inserts are very general terms. Our results give the insight of what specific disorders are actually encountered by the patients.

Newly discovered ADRs (e.g., "头晕(Dizziness)" for "易瑞沙(Iressa)"). 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.

5 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 ADRs lexicon. In order to build the evidence classifier, we propose an automatic labeling algorithm to produce large amount 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 algorithm proves to be highly effective.

Funding

This work has been supported by AstraZeneca.

Conflict of Interest: none declared.

References

Benton, A. et al., (2011) Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. J Biomed Inform., 44:989-996.

Bombardier, C. et al., (2000) Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. N Engl J Med, 343(21):1520-1528.

Bresalier, R. et al., (2005). Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. The N Engl J Med, 352 (11): 1092–1102.

Chang, P. C et al., (2009). Discriminative Reordering with Chinese Grammatical Relations Features. In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation.

Freifeld, C. C. et al., (2014) Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Saf., 37(5): 343-350.

Graham, D.J. et al., (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, Vol. 365, No. 9458, 475–481.

Gurulingappa, H. et al. (2013) Automatic detection of adverse events to predict drug label changes using text and data mining techniques. Pharmacoepidemiol Drug Saf, 22(11):1189–1194.

Hahn, U. et al., (2012) Mining the pharmacogenomics literature--a survey of the state of the art. Brief Bioinform, 13(4):460–494.

Harpaz, R. et al (2012) Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther, 91(6):1010–1021.

Harpaz, R. et al. (2010) Statistical mining of potential drug interaction adverse effects in FDA' s spontaneous reporting system. AMIA Annu Symp Proc., 281-285.

Jiang, L. et al. (2013) Discovering consumer health expressions from consumer-contributed content. SBP, 164–174.

Karimi, S. et al. (2011) what do patient forums reveal? In The second international workshop on Web science and information exchange in the medical Web. MedEX, New York, NY, USA: ACM, 10–11.

Lardon, J. et al. (2015) Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review." Ed. Gunther Eysenbach. Journal of Medical Internet Research 17.7, e171.

Leaman, R. et al. (2010)., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J. & Gonzalez, G: Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, 117–125.

Li, Y. (2011) : Medical datamining: Improving information accessibility using online patient drug reviews. PhD thesis, MIT, Dept. of Electrical Engineering and Computer Science, Cambridge, Massachusetts, USA.

Liu, M. et al. (2013) Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. J Am Med Inform Assoc, 20(3):420–426.

Liu, X and Chen, H. (2013) AZDrugMiner: An information extraction system for mining patient-reported adverse drug events in online patient forums. ICSH, 134–150.

Liu, X. et al. (2014) Identifying adverse drug events from health social media: a case study on heart disease discussion. ICSH, 25–36.

Nikfarjam, A. and Gonzalez, GH. (2011) Pattern mining for extraction ofmentions of adverse drug reactions from user comments. AMIA Annu Symp Proc., 1019-26.

Nikfarjam, A. et al. (2015) Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc, 22(3):671-81.

Sampathkumar, H. et al. (2014) Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model. BMC Med Inform Decis Mak., 14:91.

Sarker, A. and Gonzalez, G. (2015) Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform, 53:196-207.

Sarker, A. et al. (2015) Utilizing social media data for pharmacovigilance: A review, Journal of Biomedical Informatics, Volume 54, Pages 202-21

Sharif H. et al. (2014) Detecting adverse drug reactions using a sentiment classification framework. ASE SocialCom.

Sohn, S. et al., (2011) Drug side effect extraction from clinical narratives of psychiatry and psychology patients. J Am Med Inform Assoc, 18(Suppl 1):i144–i149.

Trotti, A. et al., (2003, July). CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment. In Seminars in radiation oncology (Vol. 13, No. 3, pp. 176-181). WB Saunders.

Wang, W. et al., (2011) A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations. AMIA Annu Symp Proc., 1464-1470.

Warrer, P. et al., (2012) Using text-mining techniques in electronic patient records to identify ADRs from medicine use. Br J Clin Pharmacol, 73(5):674–684.

Wu, H. et al. (2012) An early warning system for unrecognized drug side effects discovery. In Proceedings of the 21st international conference companion on WorldWide Web. New York, NY, USA: ACM, ; 2012:437–440.

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

Yang, C. C. et al. (2012) Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media. HI-KDD' 12, Beijing, China.

Yang, C. C. et al. (2012) Social mediamining for drug safety signal detection. SHB' 12, New York, NY, USA: ACM, 33–40.

Yang, C.C et al. (2012) Automatic Adverse Drug Events Detection Using Letters to the Editor. AMIA Annu Symp Proc., 1030-1039.

Ye, H. et al. (2014) Construction of drug network based on side effects and its application for drug repositioning. PloS one 9.2, e87864.

Yeleswarapu, S. et al., (2014) A pipeline to extract drug-adverse event pairs from multiple data sources. BMC Med Inform Decis Mak., 14:13.

Zhang, Hua-P. et al. (2003) HHMM-based Chinese lexical analyzer ICTCLAS. In Precedings of the second SIGHAN workshop on Chinese language processing-Volume 17 (pp. 184-187). Association for Computational Linguistics.

Appendix A List of Drugs Studied

Category	Drug Name	Manufacturer	Number of Posts	Number of Pairs
高血压(Hypertension)	缬沙坦(Valsartan)	Novartis	4729	1313
	特拉唑嗪(Terazosin)	Abbott	6101	122
	替米沙坦(Telmisartan)	Boehringer Ingelheim	2190	120
	培哚普利(Perindopril)	Servier	2470	186
	氯沙坦钾(Losartan Potassium)	Bristol-Myers Squibb	4001	423
	厄贝沙坦(Irbesartan)	Sanofi	3511	261
	吲达帕胺(Indapamide)	Servier	3600	958
	地尔硫卓(Diltiazem)	Ethypharm	3160	113
	坎地沙坦(Candesartan)	Takeda	1639	468
	富马酸比索洛尔(Bisoprolol fumarate)	Merck KGaA	6887	256
	贝尼地平盐酸盐(Benidipine	Kyowa Hakko Kirin	4277	163
	阿替洛尔(Atenolol)	AstraZeneca	5214	1555
	阿罗洛尔(Arotinolol)	Sumitomo Dainippon	2196	127
	氨氯地平(Amlodipine)	Pfizer	9674	811
	倍他乐克(Betaloc)	AstraZeneca	91477	30041
	波依定(Plendil)	AstraZeneca	23235	6462
糖尿病(Diabetes)	维格列汀(Vildagliptin)	Novartis	212	103
	捷诺维(Glactiv)	Merck & Co.	362	190
	安立泽(Onglyza)	AstraZeneca	607	250
	盐酸吡格列酮(Pioglitazone	Takeda	2048	304
	奥利司他(Orlistat)	Roche	3508	146
	甘精胰岛素(Insulin glargine)	Sanofi	7670	162
	万苏平(Glimepiride)	Sanofi	1290	104
	格列齐特(Gliclazide)	Servier	2766	238
	百泌达(Byetta)	AstraZeneca	280	194
	阿卡波糖(Acarbose)	Bayer	5287	597
哮喘(Asthma)	茶碱(Theophylline)	3M Pharmaceuticals	10298	1149
	沙丁胺醇(Salbutamol)	GlaxoSmithKline	4668	1493
	美喘清(Procaterol)	Otsuka	8864	386
	盐酸奥洛他定(Olopatadine	Kyowa Hakko Kirin	2268	641
	孟鲁司特钠(Montelukast sodium)	Merck & Co.	38857	10130
	丙酸氟替卡松(Fluticasone propionate)	GlaxoSmithKline	5840	4027
	阿米迪(Amiaid)	Nitto Denko	2519	1475
	普米克(Pulmicort)	AstraZeneca	8905	9737
	信必可(Symbicort)	AstraZeneca	8458	7553
他汀类药物(Statins)	辛伐他汀(Simvastatin)	Merck & Co.	7510	111
	瑞舒伐他汀(Rosuvastain)	AstraZeneca	5494	1952
	洛伐他汀(Lovastatin)	Merck & Co.	3871	538
乳腺癌(Breast Cancer)	阿那曲唑(Anastrozole)	AstraZeneca	37463	1823
麻醉 (Sedation)	得普利麻(Diprivan)	AstraZeneca	1097	356
胃酸过多 (GERD)	洛赛克(Omeprazole)	AstraZeneca	71525	2894
	耐信(Nexium)	AstraZeneca	69491	47025
肺癌 (Lung Cancer)	易瑞沙(Iressa)	AstraZeneca	14115	15820
鼻炎 (Rhinitis)	雷诺考特(Rhinocort)	AstraZeneca	14852	7397
精神分裂 (Schizophrenia)	思瑞康(Seroquel)	AstraZeneca	12310	10098
急性冠脉综合征 (acute coronary	倍林达(Brilinta)	AstraZeneca	430	154