|  |
| --- |
| Data and text mining  Automatic Discovery of Adverse Drug Reactions from Chinese Social Media  Quanyang Liu1\*, Meizhuo Zhang2\*, Chen Ge1, Jiemin Wang2, Jia Wei2\*\*, Kenny Q. Zhu1\*\*  1Dept. CSE, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China  2R&D Information, AstraZeneca, 199 Liangjing Road, Pudong, Shanghai, 201203, China  \*The authors contributed equally to this work. \*\*Corresponding authors |

Abstract

**Motivation:** Despite the tremendous efforts spent before every drug release, some ADRs may go undetected or unaware of and cause harm to both the users and to the pharmaceutical companies. One plausible venue to collect evidences of such ADRs is from 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 association between drug uses 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/](http://adapt.seiee.sjtu.edu.cn/~qyliu/demo-0917/) , which contains the ADRs mined for 46 popular medicines.

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

# Introduction

Determination of adverse drug reactions (ADR) is an important part of pharmaceutical research and drug development. Traditional approach has relied on extensive clinical trials and tests, which is a lengthy and expensive process. Because drug reaction is highly complex and depends on many contextual factors, despite the tremendous efforts spent before every drug release, there are still ADRs that go undetected or unaware of. Furthermore, some reactions take a long time to develop, which goes well beyond the development cycles of the drugs. For example, Vioxx, an effective medicine for arthritis and acute pain, was approved by FDA in 1999, but later was discovered to trigger heart diseases and stroke, which was previously not considered seriously in clinical trials. These unknown ADRs partly led to more than 27,000 sudden deaths around the world[[1]](#footnote-1). Even if an ADR was known in advance during clinical trial, the likelihood of its occurrence and its severity may also be ascertained during large scale commercial use itself. Discovering and understanding ADRs after commercial use is useful not only in drug research and development, but also in reducing the harm and damage to both the drug users and the drug makers.

Recent years saw 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 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 challenge because signals are embedded in natural language, which is inherently ambiguous and noisy. Biomedical literature such as scientific papers is comparatively easier to mine (Wang et al., 2011; Yang et al., 2012) because of the formal, almost error-free written language used. However, the information thereof is not up-to-date and often biased. Clinical resources were targeted by various methods, such as text mining for identifying ADRs from medicine uses (Warrer et al., 2012), rule-based method to extract side effects from clinical narratives (Sohn et al., 2011) and retrospective medication orders along with inpatient laboratory result 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 drug use experiences. The large volume, colloquial use of natural language, spelling and grammatical errors all pose significant challenges in mining ADRs from such data sources. Existing methods for social media text 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 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, it could not recognize non-regular ADRs that are not contained in the lexicon. Besides, it suffers 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 method for 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 produce reasonable accuracy. But they all need lots of training data, hence tremendous amount of manual efforts. That is also a reason why we propose a semi-supervised learning framework which requires very little manual annotations.

Although there is substantial past 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 mind ADRs from three popular social media sites, namely Xunyiwenyao[[2]](#footnote-2), Haodaifu[[3]](#footnote-3) and Sina Weibo[[4]](#footnote-4). Xunyiwenyao and Haodaifu are both online public forums for health related discussions. Each discussion thread is started with a patient’s question and followed by responses from multiple doctors or other patients (see Figure 1). Weibo is a Chinese microblogging website where a user can start a new conversation in which their friends may respond with comments or forward the discussion to other people. 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 1 Question posted on Xunyiwenyao website

And 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.

# Method

Our framework (depicted in Figure 4) is divided into four parts, namely constructing lexicons, extracting candidate ADRs, classifying evidences and finally ranking the ADRs.

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.

### Basic ADR lexicon

The basic ADR lexicon comes from three resources: Common Terminology Criteria for Adverse Events (CTCAE), Sougou Pinyin ADRs lexicon[[5]](#footnote-5) and ADR database translation[[6]](#footnote-6). CTCAE is the frequently used ADR terminology, which is also available in Chinese. Sougou ADRs is utilized particularly for colloquial terms. ADRs database coversmore than 6000 ADRs and was translated into Chinese by Google Translate[[7]](#footnote-7). In addition, classification of these terms is very important. Because some words have the same or similar meaning, their result can be merged in the following analyzing 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 two 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[[8]](#footnote-8). 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 slight variation with a degree modifier between an organ name and symptom word (such as “痛”), 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XXXXX 痛”, from online discussion corpus. The algorithm to extend ADR lexicon is presented briefly as follows.

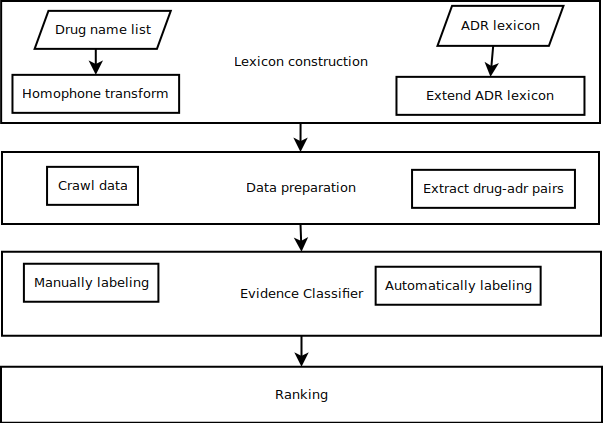


Figure 2 System framework

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

1. Data preparation

This section describes how we extract evidences of ADR for drugs from the three data sources. We discuss Weibo separately because the nature of posts on Weibo is substantially different than Xunyiwenyao and Haodaifu.

### Extraction of evidences

First, we preprocess all the user posts from three websites. If one post contains a drug name that we are interested in, 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 possible ADR terms from the effective posts. However, when a drug name X is mentioned in a post, the user may not actually take that drug. Similarly, when an ADR term is mentioned, the user may not actually have the symptom, or the symptom may not the result of taking X. So given a pair of a drug name and an ADR, discovered from a post, we need to determine whether the ADR is truly the consequence of taking the drug, given context of the pair in the post. Our context is defined as several consecutive sentences that contain a drug-ADR pair. The length of the context is a configurable parameter, a larger one would discover more pairs but may decrease the accuracy. We set it to 55 segmented units. 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, got a headache, eye diplopia and blur 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:

* A doctor would post a message after answering a question in Xunyiwenyao or Haodaifu, and which is already contained in the crawled data so it’s redundant;
* When people comment and forward a message, it rarely contains a complete sentence, which means it’s highly dependent on the original message and make it harder to processing;
* Very few messages are really about ADRs, for example, there are 7734 messages about Betaloc we crawled from Weibo, but only 1323 messages contain both Betaloc and a condition;
* There are lots of noise, such as advertisement from manufacturers, shown in Figure 3. 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.

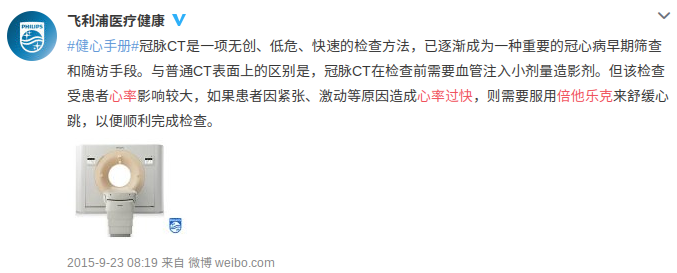


Figure 3 An advertisement on Weibo

1. Evidence Classifier

The problem of evidence classification is given a drug name and a medical condition, identified by the extended lexicon, as well as their context in the original text, 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.

### Build training set

A supervised classifier requires labeled training data. However, manual labeling on user discussion posts, which contain large amount of informal use of language and colloquial terms, is expensive and tedious. Fortunately, information in the instruction manual (package inserts) of the drugs, e.g., the indications and the known side effects of the drug, can be used to automatically generate label 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 instruction manual 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(dizzy)” 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 felt dizzy and flustered, restless, blood pressure at 160/100, tachycardia electrocardiogram 160 times, and the was given Betaloc)

And “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 doctor’s advice, Cordarone reduced to 1/4 tablets per day, plus one tablet of Betaloc (slow release). Atrial fibrillation occurs after a period of time)

Because the actual situation arising from patients experience may be more complicated than specified on the manual, we adopt a semi-supervised approach instead. We first manually label 400 sentences — 211 with positive labels and 211 with negative labels. 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 at *Features extraction* part, and it’s the same as those for our final classifier, but with more training data we can cover more unseen token. If the classifier predicts a sentence to be positive, and the medical condition is already a known ADR for the drug according to the manual, then 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 mark this sentence as a negative training instance.

With very little manual effort, we now obtain much larger number of positive and negative training data --- 17,382 training instances in total. By manual inspection, the accuracy of automatic labeling is 92%.

### Features extraction

Our main evidence classifier extracts the following the features, after parsing the evidence sentences into dependency tree:

* Verbs before the drugs, e.g. “服用(take)” in “服用倍他乐克(take Bataloc)”;
* Verbs before the conditions, e.g. “感到(feel)” in “感到头疼(feel headaches)”;
* Verbs after the conditions, e.g. “好转(improved)” in “头疼好转(headaches 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 Bataloc)”;
* Punctuation that surround drugs and conditions;
* The number of other drugs and other conditions between the drug and condition we consider;
* A Boolean value that indicates whether condition appears in front of the drug or not.

The verbs are hard to extracted without parsing the sentences, because there are always some modifiers along with them in Chinese. For example, “头疼好转(headaches improved)” would often 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.

### Train a classifier

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

1. Manually label some training data;
2. Train a classifier M’;
3. Use M’ and instruction manuals to generate more training data;
4. Train the final classifier M.
5. Ranking

For each drug, there are many possible ADRs, we are interested in ADRs of higher confidence. One way of ranking the ADRs of a drug is by the number of its appearances in positive evidences. This doesn’t work 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”. Given that the accuracy of our classifier is x%, there are still 1-x% of the mentions of “hypertension” that are incorrectly classified as ADR. 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.

# Results

We divide our evaluation into three parts. First we evaluate the accuracy of the classifier, by showing the accuracy of prediction of drug-symptom association. Then run the automatically labeling algorithm iteratively and show the change of the F1-scores. Finally, we use MRRs of ADRs and indications of instruction manuals to evaluate the end-to-end results. After these three parts, we show the top-ten discovered ADRs of several drugs, as verification and supplement for the instruction manuals.

## Data set

We have crawled user messages posted between January 2011 to April 2015 on Haodaifu and Xunyiwenyao about 46 drugs which are the treatments for 12 broadly classified diseases. Table 2 summarizes the diseases and the number of corresponding drugs. The total number of crawled posts is 456,753. After preprocessing these posts, we obtain 170,196 drug-ADR pairs.

We manually labeled a number of randomly sampled pairs (along with their contexts), and from which selected 211 positive pairs and 211 negative pairs as our initial training data.

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 |

## Drug-symptom association

As the test data set, we manually label 200 pairs of drug and ADR -- 100 positive pairs and 100 negative ones. Then we compare the classifier trained by three different labeling approaches: i) the 422 manually labeled pairs, ii) the labels generated directly from the instruction manual, and iii) semi-supervised labels. In addition to the SVM classifier, we experimented with three other baseline approaches, namely an HMM based classifier (Sampathkumar et al., 2014), a CRFs based classifier (Nikfarjam et al., 2015) and a pattern based classifier. The HMM and CRF classifiers were slightly modified to adapt to the Chinese input. For example we use ICTCLAS to segment and POS tag the input sentences. In the pattern-based classifier, we extract preposition, conjunction and noun of locality from sentences as patterns from training data generated by instruction manuals, and each pattern has a weight which is its frequency of occurrence while a negative pattern 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 match several patterns, then the score for it is the sum of these patterns. Then we build a classifier based on the score: if the score is greater than 0, it’s positive; otherwise negative. The result is shown at 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 |

## Bootstrapping of automatically labeling

Our semi-supervised automatic labeling algorithm uses the instruction manuals and 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 select the 200 manually labeled sentences mentioned above as our test set, and the 422 sentences mentioned above as our initial training data. Then run the algorithm multiple times until it converges. The result is as Figure 3.

Figure 3 Accuracy at each iteration

## End-to-end ranking

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

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 instruction manuals | **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 |

The reciprocal rank of a query reponse 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 querys.[[9]](#footnote-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.

## Top-ten discovered ADRs

We show the most frequently reported ADRs (with percentages of their occurrences in the related posts in parenthesis) for 4 different drugs for different indications in Table 5. ADRs that don’t have direct match in the instruction manuals 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) |

# Discussion

As shown at Table 3, The semi-supervised labeling approach provides the best results with F1-score significantly higher than the other approaches. The HMM based method perform worst, because it only utilizes the positive training data, so the training data is only half of the others. One interesting observation is that the positive accuracy is inversely correlated with the negative accuracy. 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.

Figure 3 shows the result of the bootstraping of our automatically labeling algorithm. We see that it will quickly converge: there are no more labeled data generated after 7th iteration. There is a dramatic improvement in accuracy from the 0th iteration to the last since more knowledge is acquired over time. The gain in accuracy saturates after a peak is reached at the 3rd iteration. We would then use the training data obtained at that time to train our final classifier.

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 patterns method, but sometimes can be worse than labeling only with instruction manuals. The reason is that only using instruction manuals the training data is not very accurate, so the trained classifier may learn some specific expressions for the instruction manuals and it can outperform our semi-supervised labeling method.

In Table 5, we discovered many ADRs that are already included in the instruction manuals. Although these ADRs are known, the frequency statistics can be valuable for: i) verifying ADRs listed in the instruction manuals; ii) studying the relative frequency between the ADRs.

There are also a number of ADRs that don’t find direct match in the manuals. These fall into several cases:

1. Synonyms of the known ADRs (eg. “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 instruction manuals are often some terminologies and the colloquial synonyms can help patients understand them easily.

1. Specialization of the known ADRs (eg. “失眠(Insomnia)” is a specialized case of “睡眠障碍(Sleep disorders)” for “耐信(Nexium)”)

Some ADRs from instruction manuals are very general terms. Our results gives the insight of what specific disorders are actually encountered by the patients.

1. Newly discovered ADRs (eg. “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.

# Conclusion

We have proposed an effective framework for extracting and analyzing ADRs from online social media. It uses lexicon-based method to extract ADRs from data, and then uses a binary classifier to find the positive evidences. In this framework, we introduce a data-driven algorithm to extend our ADRs lexicon. In order to build the evidences classifier, we propose an automatic labeling algorithm to produce large amount of labeled sentences. Completely relying on the information from the instruction manuals produces training data that is too noisy. Our tradeoff is a semin-supervised approach where we manually label a small set, then use these data and instruction manuals collectively to generate more training data. This algorithm turns out to be very effective, and it can be extended to other similar tasks.

**Acknowledgements**

Funding

This work has been supported by the …..

*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., 2011;44:989 996.

Freifeld, C. C. et al. (2014) Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Saf., 37(5): 343-50.

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.

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

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.

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.

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

Table 5 Drugs list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Categories | Drugs name | Originators | Posts | Pairs |
| 高血压(Hypertension) | 缬沙坦(Valsartan) | Novartis | 4729 | 665 |
| 压宁定(Urapidil) | Takeda | 8403 | 19 |
| 特拉唑嗪(Terazosin) | Abbott | 6101 | 65 |
| 替米沙坦(Telmisartan) | Boehringer Ingelheim | 2190 | 61 |
| 瑞泰(Ramipril) | Sanofi | 1092 | 6 |
| 培哚普利(Perindopril) | Servier | 2470 | 68 |
| 氯沙坦钾(Losartan Potassium) | Bristol-Myers Squibb | 4001 | 187 |
| 氯沙坦钾-氢氯噻嗪(losartan/hydrochlorothiazide) | Bristol-Myers Squibb | 1803 | 12 |
| 赖诺普利(Lisinopril) | Merck & Co. | 1646 | 23 |
| 乐卡地平(Lercanidipine) | Recordati | 268 | 0 |
| 拉西地平(Lacidipine) | GlaxoSmithKline | 1539 | 24 |
| 伊索格拉啶(Irsogladine) | Nippon Shinyaku | 527 | 4 |
| 厄贝沙坦(Irbesartan) | Sanofi | 3511 | 127 |
| 吲达帕胺(Indapamide) | Servier | 3600 | 501 |
| 咪达普利(Imidapril) | Mitsubishi Tanabe Pharma | 1717 | 22 |
| 福辛普利(Fosinopril) | Bristol-Myers Squibb | 1876 | 0 |
| 多沙唑嗪(Doxazosin) | Pfizer | 1207 | 13 |
| 地尔硫卓(Diltiazem) | Ethypharm | 3160 | 41 |
| 地拉普利(Delapril) | Takeda | 95 | 0 |
| 卡维地洛(Carvedilol) | Roche | 724 | 42 |
| 坎地沙坦(Candesartan) | Takeda | 1639 | 0 |
| 布那唑嗪(Bunazosin) | Eisai | 2 | 0 |
| 富马酸比索洛尔(Bisoprolol fumarate) | Merck KGaA | 6887 | 95 |
| 贝尼地平盐酸盐(Benidipine hydrochloride) | Kyowa Hakko Kirin | 4277 | 77 |
| 阿替洛尔(Atenolol) | AstraZeneca | 5214 | 1555 |
| 尼群地平(Nitrendipine) | Bayer | 1329 | 11 |
| 阿罗洛尔(Arotinolol) | Sumitomo Dainippon Pharma | 2196 | 54 |
| 氨氯地平(Amlodipine) | Pfizer | 9674 | 385 |
| 阿利沙坦酯(Allisartan isoproxil) | Allist Pharmaceuticals | 3 | 0 |
| 阿利吉仑(Aliskiren) | Novartis | 130 | 0 |
| 倍他乐克(Betaloc) | AstraZeneca | 91477 | 14045 |
| 波依定(Plendil) | AstraZeneca | 23235 | 3340 |
| 糖尿病(Diabetes) | 倍欣(Voglibose) | Takeda | 1200 | 11 |
| 维格列汀(Vildagliptin) | Novartis | 212 | 0 |
| 捷诺维(Glactiv) | Merck & Co. | 362 | 60 |
| 安立泽(Onglyza) | AstraZeneca | 607 | 15 |
| 马来酸罗格列酮片(Rosiglitazone maleate) | GlaxoSmithKline | 1219 | 19 |
| 孚来迪片(Repaglinide) | Boehringer Ingelheim | 2054 | 23 |
| 盐酸匹格列酮(Pioglitazone hydrochloride) | Takeda | 2048 | 156 |
| 奥利司他(Orlistat) | Roche | 3508 | 69 |
| 那格列奈(Nateglinide) | Ajinomoto | 2268 | 27 |
| 米格列奈钙片(Mitiglinide) | Kissei | 32 | 0 |
| 瑞格列奈二甲双胍组合物(Metformin repaglinide) | Novo Nordisk | 3 | 0 |
| 诺和力(Liraglutide) | Novo Nordisk | 168 | 0 |
| 长秀霖(Basalin) | Gan & Lee | 983 | 40 |
| 甘精胰岛素(Insulin glargine) | Sanofi | 7670 | 73 |
| 门冬胰岛素注射液(Insulin Aspart) | Novo Nordisk | 1894 | 0 |
| 格列吡嗪控释片(Glipizide extended-release) | Pfizer | 525 | 1 |
| 万苏平(Glimepiride) | Sanofi | 1290 | 0 |
| 格列齐特(Gliclazide) | Servier | 2766 | 126 |
| 百泌达(Byetta) | AstraZeneca | 280 | 0 |
| 利拉利汀(Linagliptin) | Boehringer Ingelheim | 98 | 0 |
| 苯甲酸阿格列汀(Alogliptin benzoate) | Takeda | 28 | 0 |
| 阿卡波糖(Acarbose) | Bayer | 5287 | 265 |
| 哮喘(Asthma) | 扎鲁司特(Zafirlukast) | AstraZeneca | 135 | 0 |
| 茶碱(Theophylline) | 3M Pharmaceuticals | 10298 | 230 |
| 沙丁胺醇(Salbutamol) | GlaxoSmithKline | 4668 | 483 |
| 美喘清(Procaterol) | Otsuka | 8864 | 102 |
| 吡嘧司特钾(Pemirolast potassium) | Bristol-Myers Squibb | 2290 | 0 |
| 盐酸奥洛他定(Olopatadine hydrochloride) | Kyowa Hakko Kirin | 2268 | 0 |
| 孟鲁司特钠(Montelukast sodium) | Merck & Co. | 38857 | 1873 |
| 洛草氨酸氨丁三醇(Lodoxamide tromethamine) | Pfizer | 1856 | 0 |
| 福莫特罗(Formoterol) | AstraZeneca | 1613 | 33 |
| 丙酸氟替卡松(Fluticasone propionate) | GlaxoSmithKline | 5840 | 715 |
| 依匹斯汀(Epinastine) | Boehringer Ingelheim | 2163 | 0 |
| 阿米迪(Amiaid) | Nitto Denko | 2519 | 433 |
| 盐酸班布特罗(Bambuterol hydrochloride) | AstraZeneca | 952 | 16 |
| 普米克(Pulmicort) | AstraZeneca | 8905 | 3984 |
| 信必可(Symbicort) | AstraZeneca | 8458 | 3484 |
| 他汀类药物(Statins) | 阿伐他汀(Atorvastatin) | Pfizer | 2464 | 0 |
| 辛伐他汀(Simvastatin) | Merck & Co. | 7510 | 67 |
| 瑞舒伐他汀(Rosuvastain) | AstraZeneca | 5494 | 1202 |
| 普伐他汀(Pravastatin) | Sankyo Pharma Inc. | 2204 | 10 |
| 洛伐他汀(Lovastatin) | Merck & Co. | 3871 | 243 |
| 氟伐他汀(Fluvastatin) | Novartis | 2980 | 22 |
| 依折麦布辛伐他汀(Ezetimibe simvastatin) | Merck & Co. | 1557 | 0 |
| 匹伐他汀(Pitavastatin) | Nissan Chemical Industries | 105 | 0 |
| 西立伐他汀(Cerivastatin) | Pfizer | 43 | 0 |
| 美伐他汀(Mevastatin) | Daiichi-Sankyo | 29 | 0 |
| 氨氯地平阿托伐他汀(Atorvastatin amlodipine) | Pfizer | 209 | 0 |
| 烟酸辛伐他汀(Niacin simvastatin) | Abbott Laboratories | 1 | 0 |
| 烟酸洛伐他汀(Niacin lovastatin) | KOS Pharmaceuticals, Inc | 0 | 0 |
| 西格列汀辛伐他汀复合剂(Simvastatin sitagliptin) | Merck & Co. | 0 | 0 |
| 乳腺癌 | 阿那曲唑(Anastrozole) | AstraZeneca | 37463 | 1074 |
| 麻醉 | 得普利麻(Diprivan) | AstraZeneca | 1097 | 159 |
| 胃酸过多 | 洛赛克(Omeprazole) | AstraZeneca | 71525 | 1248 |
| 肺癌 | 易瑞沙(Iressa) | AstraZeneca | 14115 | 10028 |
| 胃病 | 耐信(Nexium) | AstraZeneca | 69491 | 29754 |
| 鼻炎 | 雷诺考特(Rhinocort) | AstraZeneca | 14852 | 1419 |
| 精神分裂 | 思瑞康(Seroquel) | AstraZeneca | 12310 | 4064 |
| 急性冠脉综合征 | 倍林达(Brilinta) | AstraZeneca | 430 | 67 |

1. https://en.wikipedia.org/wiki/Rofecoxib [↑](#footnote-ref-1)
2. http://club.xywy.com/ [↑](#footnote-ref-2)
3. http://www.haodf.com/ [↑](#footnote-ref-3)
4. http://weibo.com [↑](#footnote-ref-4)
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> . [↑](#footnote-ref-5)
6. A English database comes from Ye et al., 2014. [↑](#footnote-ref-6)
7. https://translate.google.com/ [↑](#footnote-ref-7)
8. http://www.meddra.org/ [↑](#footnote-ref-8)
9. https://en.wikipedia.org/wiki/Mean\_reciprocal\_rank [↑](#footnote-ref-9)