SideEffectPTM: An Unsupervised Topic Model to Mine Adverse Drug Reactions from Health Forums

Sheng Wang¹, Yanen Li¹, Duncan Ferguson², Chengxiang Zhai¹

¹Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{swang141,yanenli2,czai}@illinois.edu

²Department of Comparative Biosciences, University of Illinois at Urbana-Champaign, Urbana, IL 61802, USA
dcf@illinois.edu

ABSTRACT
Automatic discovery of medical knowledge using data mining has great potential benefit in improving population health and reducing healthcare cost. Discovering adverse drug reaction (ADR) is especially important because of the significant morbidity of ADRs to patients. Recently, more and more patients describe the ADRs they experienced and seek for help through online health forums, creating great opportunities for these forums to discover previously unknown ADRs.

In this paper, we propose a novel unsupervised approach to tap into the increasingly available health forums to mine the side effect symptoms of drugs mentioned by forum users. Our approach is based on a novel probabilistic mixture model of symptoms, where the side effect symptoms and disease symptoms are explicitly modeled with two separate component models, and discovery of side effect symptoms can be achieved in an unsupervised way through fitting the mixture model to the forum data. Extensive experiments on online health forums demonstrate that our proposed model is effective for discovering the reported ADRs on forums in a completely unsupervised way. The mined knowledge using our model is directly useful for increasing our understanding of more challenging ADRs, such as long-term side effects, drug-drug interactions, and rare side effects. Since our approach is unsupervised, it can be applied to mining large amounts of growing forum data to discover new knowledge about ADRs, helping many patients become aware of possible ADRs.

Categories and Subject Descriptors
J.3 [Life and Medical Sciences]: Life and Medical Sciences—Health

General Terms
Algorithms, Performance, Experimentation

Keywords
Adverse Drug Reaction, Probabilistic Topic Model, Health Forum

1. INTRODUCTION
Automatic discovery of medical knowledge using data mining has great potential benefit in improving population health and reducing healthcare cost. Applying data mining to discover adverse drug reactions (ADR) [6, 18, 20, 21, 34, 37] is especially important because of the significant morbidity of ADRs to patients. An adverse drug reaction refers to the harm associated with the use of given medication [24]. It is the fourth leading cause of death in the U.S., following heart disease, cancer and stroke [13]. A recent survey shows that ADRs account for 5% of all hospital admissions, occur in 10-20% of hospital inpatients and cause deaths in 0.1% of medical and 0.01% of surgical inpatients [1, 2, 35]. Approximately $136 billion has been spent on treating ADRs each year in the U.S. [19, 31]. Besides these shocking statistics, ADRs also affect a patient’s quality of life, reduce confidence in physicians and increase patient care costs [26]. ADRs would also mimic disease, resulting in unnecessary investigations and delay in treatment [26].

Although clinical trials are used to examine drug safety before drugs reach the market, they can hardly uncover all ADRs due to the limited patient-year experience. The post marketing drug safety mainly relies on spontaneous reports from hospitals, pharmaceutical companies and patients to the Adverse Event Reporting System (AERS). However, it may take years for The Food and Drug Administrations (FDA) to withdraw or restrict dangerous drugs based on these reports. For example, Trovafloxacin, which went on the market in Feb 1998 as a broad spectrum antibiotic, was not withdrawn until June 1999 due to the risk of hepatotoxicity. The drug producers are thus seeking a shorted timeline to detecting the potential danger of their new drugs. Patients and doctors also need more informative and detailed examples of cases about ADRs instead of a mere drug-symptom pair. In addition, the AERS is often difficult for patients to use. The significant mismatches between patient terminology and both the information source terminology and standard medical vocabularies often lead to confusion and misunderstanding [21, 38]. In this paper, we explore a promising al-
ternative way of discovering ADRs by mining the growing online health forum data where many patients voluntarily report detailed cases about suspicious ADRs.

Recently, the dramatic growth of social networks has changed our lives in many ways, especially in how people interact with each other and how people obtain information. In particular, as ad hoc social networks, online health forums have attracted more and more patients to describe the ADRs they experienced and seek for help [6, 22, 36, 37]. Online health forums contain the richer detailed text information about patients’ ADRs as well as other patients’ similar experience in comparison to general social networks [18]. The wide coverage of drugs and abundant patients’ personal experience make online health forum a clearly valuable source for mining previously unknown ADRs. In this paper, we study how to mine ADRs from online health forums. Adding to the current practice of AERS reporting, this new way of discovering ADRs can potentially allow us to discover ADRs much faster, and may also be advantageous in discovering long-term or rare ADRs and drug-drug interactions. The goal of our ADRs mining is to improve sensitivity to detect new drug safety signals while the goal of the FDA based AERS is to distinguish the drug action from other factors including the disease treated. In addition, we can link the mined ADRs to the related threads in health forums to provide richer case descriptions to users in order to complement the brief information from AERS.

![Figure 1: An example of a forum thread. The red rectangles refer to the drugs. The green ellipses refer to the disease symptoms. The blue lines refer to the side effect symptoms.](image)

There have been several previous studies on discovering ADRs from health forums [6, 16, 22, 36, 37]. However, most of these approaches are based on supervised machine learning, which requires labeled data. Unfortunately, annotating ADRs is not only time consuming but also requiring professional medical background. This limits their utility, and prevents them from being applied in large scale. In [16], an unsupervised approach was proposed to organize and integrate patient outcomes to different drugs. However, it only discovers opinion topics to the drugs, but cannot generate symptoms describing ADRs. In this paper, we propose a novel unsupervised approach to tap into the increasingly available health forums to mine the side effect symptoms of drugs mentioned by forum users. Modeling the free text in online health forums to discover ADRs is a challenging task. First, in order to help others better understand their conditions, patients tend to provide rather detailed descriptions, which contain both the symptoms treatable by the drugs as well as the side effect symptoms caused by the drugs. In this paper, we named the former symptoms as disease symptoms and the latter as side effect symptoms. Figure 1 is an example of a thread which contains both side effect symptoms and disease symptoms. The mix of these symptoms makes mining side effect symptoms more challenging. To the best of our knowledge, none of the existing approach explicitly separates side effect symptoms and disease symptoms when mining ADRs. Second, in a forum thread, patients may describe more than one drug while the symptom and drug name are not restricted to be in the same sentence. This increases our difficulties in modeling the free text. Therefore, it is important for us to have the potential ability to mine ADRs in any natural language.

Our approach overcomes deficiencies of previous works by a novel probabilistic mixture model of symptoms, where the side effect symptoms and disease symptoms are explicitly modeled by two separate component models and discovery of side effect symptoms can be achieved in an unsupervised way through fitting the mixture model to the forum data. Specifically, we cast the disease symptom topics as a prior in the probabilistic topic model and continue to update them when we obtain new threads. We fit the model to the text collection with the Maximum A Posterior (MAP) estimation. With the help of prior, we can naturally obtain both the disease symptoms and side effect symptoms simultaneously.

Extensive experiments on online health forums demonstrate that our proposed model is effective for discovering the reported ADRs on forums in a completely unsupervised way. We verify our model through both qualitative and quantitative evaluation on real world online health forums. We successfully find many ADRs such as “high blood sugar” to Ativan(R). The mined knowledge using our model is directly useful for enriching our knowledge of long-term side effects, drug-drug interactions, and rare side effects. We discover that Cipro(TM) may cause long-term side effect of abdominal pain. Since our approach is completely unsupervised, it can be applied to mining large amounts of growing forum data in any natural language to discover new knowledge about ADRs, helping many patients become aware of possible ADRs.

Because the FDA database is hard to use and slow in reacting to patients’ reports, our approach can complement it by providing timely and detailed cases about unknown ADRs of some new drugs to drug producers. Meanwhile, patients and doctors can only acquire information about the dangerous drug and its related ADRs through the FDA database. If they want to obtain more helpful detailed personal experience rather than the brief information from FDA, our approach can help them link the ADRs to the related documents in health forums. Moreover, the related documents are not restricted to be threads in health forums. Any document (e.g., medical literatures, news articles) can be mined by our model to provide detailed information to users.

2. RELATED WORK

Data mining is becoming a critical tool for healthcare. For example, Prather et al. [27] applied data mining approaches to search for relationships in a large clinical database. Recently, Songdi et al. [29] proposed the SympGraph to mine clinical notes through symptom relation graphs. Ryen et al. [33] presented a longitudinal log-based study of medical search and browsing behavior.
Data mining has been applied to various types of data for mining ADRs. Researchers have identified new drug safety signals from the international database of more than 2 million cases reports upon experts from around the world [20]. There are some existing works focused on text data. Liu et al. [21] extracted a wide range of disorders from online patient-submitted drug reviews. Their text data was semi-structured where patients explicitly wrote the side effect symptoms in the specification. As a result, their approach could not be applied to our completely unstructured data. Friedman extracted side effect information from clinical notes [12]. However, clinical notes were highly regularized with fewer grammar and spelling mistakes in comparison to health forums. Leaman et al. [18] mined ADRs from short messages in social networks. Their text data was shorter and cleaner in comparison to health forums. We compared our method with their approach in our experiment.

Some existing works aim at mining ADRs from online health forums. Jiang et al. [16] organized and integrated patient outcomes to different drugs through unsupervised learning. They mined different topics for the drug instead of ADRs for that drug. Chee et al. [6] identified dangerous drugs based on messages discussed in the health forums through supervised learning. We not only predicted dangerous drugs but also provided the most likely side effect symptoms for the given drug. In addition, our approach was unsupervised, which didn’t require the expensive labeled data. Yates et al. extracted different kinds of features and used conditional random field to extract ADRs. Their approach was supervised and could not distinguish the disease symptoms from the side effect symptoms [36, 37]. Liu and Chen proposed a statistical learning based approach to mine ADRs [22]. Their approach had the same limitation as [36, 37]. None of the existing approaches explicitly distinguish the side effect symptoms from the disease symptoms which often appear in the same thread. We demonstrate the importance of separating these two kinds of symptoms in our experiment.

Recently, topic model has attracted increasing attention. It has been widely and successfully applied to forum data and other text collections to mine topic patterns [3, 14, 28]. Our work contributes yet another novel use of such models for mining ADRs. Following the similar idea of defining prior [23], we define a prior to distinguish side effect symptom topics from disease symptom topics.

3. PROBLEM DEFINITION

In this section, we define the problem of mining adverse drug reactions from health forums and give the notations.

Given a collection of text threads from the health forums, our goal is to mine the ADRs for all the drugs mentioned in the text threads. The input of the mining problem is the collection of all forum threads. The output is the side effect symptoms for all the drugs mentioned in the text threads. This problem definition is different from the formulation used in existing supervised approaches where the input also includes labeled training data, which is expensive to create. We do not assume the availability of labeled data, and attempt to solve the problem in a completely unsupervised way, which would make it possible to mine all the health forum data to discover ADRs.

Each thread in the health forums can be generally regarded as a free-style document \(d\). In this paper, we refer to forum thread as document. The collection of all the documents can be denoted as \(D_0 = \{d_1, ..., d_n\}\). To leverage existing resources for ADR discovery, we assume that we have available drug and symptom recognizers. Specifically, we can use [4] to recognize drugs and symptoms. In this paper, we focus on developing statistical approaches for discovering ADRs, thus we do not explore how to improve recognition accuracy, which would be an orthogonal task.

All the drugs mentioned in the collection can be denoted as \(M = \{m_1, ..., m_p\}\), where \(m_i\) is a drug. All the symptoms mentioned in the collection can be denoted as \(S = \{s_1, ..., s_q\}\), where \(s_i\) is a symptom.

An online health forum is different from drug review system in that it doesn’t contain specific fields (e.g., symptom, drug) for patients to input. Furthermore, patients could describe more than one drug in one document in order to report the drug-drug interactions or compare different drugs. In addition, patients may describe more than one symptom they experienced. The symptom and the corresponding drug are also not necessarily in the same sentence. Therefore, the document is completely unstructured text which makes it nontrivial to connect the symptoms with the drugs.

Formally, for a given document \(d\), we denote its drug set by \(M_d = \{m_1, ..., m_{|M_d|}\}\) where \(m_i\) is a drug mentioned in the document \(d\). For a given document \(d\), we denote its symptom set by \(S_d = \{s_1, ..., s_{|S_d|}\}\) where \(s_i\) is a symptom described in the document \(d\).

In the document \(d\), not all the mentioned symptoms are necessary to be side effect symptoms. As we can see from the example in Figure 1, to describe their conditions clearly, lots of patients not only mention the side effect symptoms but also describe the disease symptoms. In this paper, side effect symptoms refer to the adverse drug reactions coming from that drug while disease symptoms refer to the symptoms that a patient wants to treat by using that drug. For example, a patient who suffered from depression would likely describe the depression symptoms (e.g., downswings in mood, insomnia) he/she had. The patient would also describe the unexpected side effect symptoms (e.g., nausea, black spot in vision) after using an anti-depressant drug. Therefore, to mine the ADRs correctly, it is necessary to further distinguish the side effect symptoms from the disease symptoms. For the given symptom set \(S_m\) of drug \(m\), we denote the side effect symptom set as \(R_m = \{r_1, ..., r_{|R_m|}\}\) where \(r_i\) is a side effect symptom of drug \(m\). We denote the disease symptom set as \(T_m = \{t_1, ..., t_{|T_m|}\}\) where \(t_i\) is a disease symptom of drug \(m\). For these three sets, we have the following equations:

\[
S_m = R_m \cup T_m, \quad R_m \cap T_m = \emptyset.
\]

To the best of our knowledge, none of the existing work has attempted to model \(T_m\) and \(R_m\) separately when mining ADRs from health forums. This would cause these approaches to misidentify the disease symptom as a side effect symptom, making the extracted ADRs noisy and incorrect. It is even more challenging if we want to model these two kinds of symptoms simultaneously through unsupervised learning since we don’t have manually labeled data to help us separate them. In our problem definition, we formally separate these two kinds of symptoms and explicitly model them simultaneously through our proposed unsupervised model. In this way, our model achieves better results in mining ADRs even without expensive human annotation.
4. SIDE EFFECT PROBABILISTIC TOPIC MODEL

One natural approach to solving our problem is to design a rule-based information extraction approach which is undesirable due to the huge amount of irregular unstructured text in health forums. In this paper, we propose to solve this task through unsupervised probabilistic topic model [3, 14] which is not only unsupervised but has also been widely applied to many of text mining problems with promising results [23, 28]. We propose to use unsupervised topic model to mine the ADRs because of the following reasons. Firstly, this is a completely unsupervised approach that does not need human annotations. As we have introduced before, labeling ADRs not only is time consuming but also requires a medical background. Secondly, topic models can deal with the documents containing more than one drug through grouping the symptoms of the same drug together and separating the symptoms of different drugs from each other. Finally, with the help of “bag of words” assumption in topic model, drug and related symptoms are no longer restricted to be in the same sentence. Our model is the first attempt to use topic model to mine ADRs from health forums.

Probabilistic topic model maps words into lower dimensional topics. It groups the words that co-occurred frequently into same topic and separates words co-occurred infrequently into different topics. Each topic is in the form of a multinomial word distribution (e.g., unigram language model). For example, a topic about a disease like “depression”, may have high probabilities for words such as “depressed” and “sleepless.” We then assume the text data to be a sample of words drawn from a mixture of many such word distributions each characterizing a topic. By fitting such a mixture model to the text data, we can find what multinomial word distributions best explain our data, and such discovered word distributions can be regarded as the topics mined from the data set.

Many of the side effect symptom expressions are in the form of phrases instead of words (e.g. tissue damage, weight gain). At the same time, these phrases are often meaningless or ambiguous when we break them into words. In contrast to conventional topic models which often process the text in word level [14, 23], we model the text in phrase level to better capture and reveal the exact symptoms. For convenience, in the rest of the paper, we will sometimes use “words” to refer to such phrase units.

Our starting point is to use a basic topic model PLSA [14] which is able to mine topics from the text collection. In PLSA, each word \( w \) in \( W \) in the document \( d \) in \( D_0 \) is associated with an unobserved class variable \( z \) in \( Z = \{z_1, ..., z_n\} \) which represents the underlying topic it belongs to. A joint probability over the observed co-occurrence table \( D \times W \) is defined by the mixture model:

\[
P(w, d) = \sum_{z \in Z} P(z) P(w|z) P(d|z).
\]

Under this definition, the log likelihood function of the collection \( D_0 \) is

\[
\log P(D_0|\Lambda) = \sum_{d \in D_0} \sum_{w \in W} c(w, d) \times \log p(w, d),
\]

where \( \Lambda \) is the set of all model parameters, \( W \) is the set of all the words and \( c(w, d) \) is the count of word \( w \) in document \( d \).

Our main idea is to design one topic to capture ADRs for each drug. Therefore, we can define a multinomial word distribution which describes the most likely ADRs for each drug. Patients tend to discuss both side effect symptoms and disease symptoms in the same document. As a result, these two kinds of symptoms may be mixed together. If we just fit the data with a topic model including only one multinomial word distribution for each drug, we would be unable to separate these two kinds of symptoms and misidentify the disease symptom as a side effect symptom. In order to separate the disease symptoms and the side effect symptoms, we assign two topics explicitly to each drug, one is the disease symptom topic and the other is the side effect symptom topic. The disease symptom topic provides the probability that a word is a disease symptom for the given drug. The side effect symptom topic provides the probability that a word is a side effect symptom for the given drug. We thus have \( 2|M| \) total topics.

We further assume that each document can only generate topics from the drugs it has mentioned. This assumption is reasonable because if a document contains a drug, it is unlikely that the patient would discuss side effects of a different drug other than the drug mentioned in the document. In this way, the topics of each document can be more concise and the words of each topic can be more discriminative. Formally, we have

\[
\forall m \notin M_d, P(T_m|d) = 0, P(R_m|d) = 0,
\]

\[
\sum_{m \in M_d} \{P(T_m|d) + P(R_m|d)\} = 1.
\]

Without ambiguity, we denote \( p(w|T_m) \) as the disease symptom topic of drug \( m \) and \( p(w|R_m) \) as the side effect symptom topic of drug \( m \). The probability of observing word \( w \) in document \( d \) is thus:

\[
P(w, d) = \sum_{m \in M_d} P(R_m)P(w|R_m)P(d|R_m)
+ \sum_{m \in M_d} P(T_m)P(w|T_m)P(d|T_m).
\]

Because the document is free-style plain English text, it could contain many background words. For example, words such as “doctor” and “help” occur frequently in the document when patients describe their conditions. However these words have little connection with the disease symptoms and the side effect symptoms, we assign two topics explicitly to each drug other than the drug mentioned in the document. As a result, these two kinds of symptoms may be mixed together. If we just fit the data with a topic model including only one multinomial word distribution for each drug, we would be unable to separate these two kinds of symptoms and misidentify the disease symptom as a side effect symptom. In order to separate the disease symptoms and the side effect symptoms, we assign two topics explicitly to each drug, one is the disease symptom topic and the other is the side effect symptom topic. The disease symptom topic provides the probability that a word is a disease symptom for the given drug. The side effect symptom topic provides the probability that a word is a side effect symptom for the given drug. We thus have \( 2|M| \) total topics.

We further assume that each document can only generate topics from the drugs it has mentioned. This assumption is reasonable because if a document contains a drug, it is unlikely that the patient would discuss side effects of a different drug other than the drug mentioned in the document. In this way, the topics of each document can be more concise and the words of each topic can be more discriminative. Formally, we have

\[
\forall m \notin M_d, P(T_m|d) = 0, P(R_m|d) = 0,
\]

\[
\sum_{m \in M_d} \{P(T_m|d) + P(R_m|d)\} = 1.
\]

Without ambiguity, we denote \( p(w|T_m) \) as the disease symptom topic of drug \( m \) and \( p(w|R_m) \) as the side effect symptom topic of drug \( m \). The probability of observing word \( w \) in document \( d \) is thus:

\[
P(w, d) = \sum_{m \in M_d} P(R_m)P(w|R_m)P(d|R_m)
+ \sum_{m \in M_d} P(T_m)P(w|T_m)P(d|T_m).
\]
a word \( w \) in document \( d \), we first flip a coin to decide whether to talk about a symptom. With probability \( \lambda_d \), we will choose to talk about a symptom rather than a background word. If so, we will further choose a drug to talk about and decide whether to talk about a disease symptom or side effect symptom for this drug. If we are talking about a side effect symptom, we will draw a word according to \( P(w|T_m) \). Otherwise, we are talking about a disease symptom and we will draw a word according to \( P(w|R_m) \). If we are talking about a background word, we will draw a word according to \( P(w|B_d) \).

From Equation 6, we can see that our model has the following parameters: \( P(w|T_m) \), \( P(w|R_m) \), \( P(d|T_m) \), \( P(d|R_m) \), \( P(T_m) \) and \( P(R_m) \). We denote all of them by \( \Lambda \). We can estimate these parameters by using maximum likelihood estimator

\[
\hat{\Lambda} = \arg\max_{\Lambda} P(D_0|\Lambda). \tag{7}
\]

We can use expectation maximization algorithm to perform maximize likelihood estimation of this latent variable model. EM alternatively performs E-step and M-step. Since the parameter estimation for side effect symptom topics is similar to disease symptom topics, we only give the update formula of side effect symptom topics due to the space limitation. In the E-step, we estimate the posterior probabilities for \( P(R_m|d, w) \).

\[
P(R_m|d, w) \propto \lambda_B P^\alpha(R_m) P^\alpha(w|R_m) P^\alpha(d|R_m) \tag{8}
\]

In the M-step, we update the parameters based on the posterior probabilities obtained in the E-step.

\[
P^{n+1}(w|R_m) \propto \sum_{d \in D_0} c(w, d) P(R_m|d, w) \tag{9}
\]

\[
P^{n+1}(d|R_m) \propto \sum_{w \in W} c(w, d) P(R_m|d, w) \tag{10}
\]

\[
P^{n+1}(R_m) \propto \sum_{w \in W} \sum_{d \in D_0} c(w, d) P(R_m|d, w) \tag{11}
\]

We could have directly applied the above model to mining side effect symptoms from the forum collection. However, the resultant topic cannot be well aligned to the side effect symptom topic and the disease symptom topic. In order to ensure the alignment, we need to force one of the topics of each drug to be aligned to the disease symptoms, thus leading the other topic describing the side effect symptoms. In probabilistic models, this can be achieved by introducing a conjugate prior into the model and using the Maximum A Posteriori estimator instead of the maximum likelihood estimator. Since we aim at mining unknown side effect symptoms from the text collection, we don’t incorporate any prior for the side effect symptom topic.

5. APPLICATION OF PROPOSED MODEL

In this section, we show how to use our model to obtain the ADRs of each drug. In addition, we show how to naturally mine more challenging side effects and link the ADRs to related documents.

5.1 Mining Side Effect Symptoms and Disease Symptoms

After we finish parameter estimation, \( p(w|R_m) \) would give us a distribution over all the phrases and those with the highest probabilities can be taken as potential ADRs of drug \( m \). Note that, our phrase dictionary not only contains symptom but also contains other background phrases such as "high dose" and "small pill". As discussed in Section 4, we try to use a background topic to decrease background phrases’ probability in the final \( p(w|R_m) \).

5.2 Mining More Challenging Side Effects

According to \([30]\), long-term side effects, rare side effects and drug-drug interactions are three kinds of more challenging ADRs in medical domain. The extreme difficulties in detecting such ADRs makes them more dangerous to patients.

- Long-term Side Effects
A long-term side effect refers to the side effect occurring after patients use the drug for a long time. This kind of side effect is hard to detect in the limited clinical trials period. We mine long-term side effects in the following steps. We first match the documents containing the time keywords such as "years" and "long time". We then extract the candidate symptom and drug pairs from these documents. Finally, we compute the likelihood of a symptom $s$ to be the long-term side effect of drug $m$ as:

$$LT(s, m) = freq(s, m) \{ P(s|R_m) - P(s|T_m) \},$$  \hspace{1cm} (15)

where $freq(s, m)$ is the frequency that $s$ and $m$ appear together within a document containing a time keyword. $P(s|R_m) - P(s|T_m)$ indicates how likely $s$ is a side effect symptom of $m$ instead of a disease symptom. Therefore, the larger the score, the more likely it would be a long-term side effect.

- **Drug-drug interactions**

Drug-drug interaction refers to the side effects appearing when using two drugs together. We extract potential dangerous drug pairs $m_i$ and $m_j$ based on the following equation:

$$DD(m_i, m_j) = P(m_i|R_{m_j}) \frac{freq(m_i, m_j)}{\sum_{m_k \in M} freq(m_i, m_k)}. \hspace{1cm} (16)$$

We regard $m_i$ as a phrase and compute the probability that it is generated from the side effect symptom topic of Drug $m_j$. This probability is then multiplied by the fraction that $m_i$ and $m_j$ appear together to get the final score. Note that, this equation is not symmetrical so that we need to consider both $DD(m_i, m_j)$ and $DD(m_j, m_i)$.

- **Rare Side Effects**

Rare side effect often refers to the side effect that between one in 1,000 and one in 10,000 people are affected. Unfortunately, we don’t have enough power in this study to support the mining of rare side effects. As a result, we focus on another type of rare side effects which are the side effects that only occur among very few drugs. If patients were affected this kind of rare side effects, our algorithm would likely be able to connect it with that drug. Formally, we define the probability that a side effect symptom $s$ is a rare side effect of drug $m$ as:

$$P(s)_m = \frac{P(s|R_m)}{\sum_{m' \in M} P(s|R_{m'})}. \hspace{1cm} (17)$$

Our intuition is similar to the idea of inverse document frequency in information retrieval. We divide the probability that the symptom is generated by the given drug by the probability that the symptom is generated by any drug.

We show the results of mining these three kinds of challenging ADRs in Section 6.4.

5.3 Linking ADRs to Related Documents

The FDA database only provides the information of dangerous drugs with their related ADRs. However, drug producers, patients and doctors often need more detailed patient cases to better analyze these unknown ADRs or compare these cases with the conditions they met. Therefore, it is important to link the extracted ADRs with the related threads and show them to patients. We link the document with the extracted ADRs through the following equations:

$$P(d)_s = p(m|d) \cdot \frac{freq(s, d)}{|d|}. \hspace{1cm} (18)$$

The extracted ADRs would be linked to the document that has a larger $p(m|d)$ and more $s$. Our model is not restricted to the forum threads and can be applied to any free text. Therefore, we can leverage the related documents retrieved from the search engine as collection and then use our model to retrieve the most related documents.

6. EXPERIMENTAL RESULTS

In this section, we perform extensive experiments on real-world online health forum to evaluate the effectiveness of our model. We perform both qualitative evaluation and quantitative evaluation. In the qualitative evaluation, we show the top side effect symptoms we have mined. In the quantitative evaluation, we compare our model with baselines on two ground truth labeling sets. We further use our model to mine some more challenging ADRs, such as long-term side effects, drug-drug interactions and rare side effects. Finally, we give a detailed comparison between the side effect symptoms we have mined and the side effect symptoms reported in the FDA database. We show that our model has mined some suspicious ADRs which have not caught enough attention by FDA.

6.1 Data Collection and Preprocessing

We crawl data from a real-world online health forum HealthBoards (www.healthboards.com). HealthBoards is an online health forum that allows patients to discuss their conditions. It has been rated as one of the top 20 health information communities, with over 10 million monthly visitors, 850,000 registered members and over 4.5 million messages posted. We crawl a large text collection of 330,305 threads. We then extract 886 threads from a board called "Drug Interaction/Side Effects." Existing NLP tools perform reasonable well when we extract drugs from the raw text. We use Metamap [4] to extract 287 drugs from all the 886 documents. Each thread has on average 336 words. As we have mentioned before, each thread may describe more than one drug. In our data, we find an average of 3.18 drugs per thread. We filter the stopwords in the text and use a large medical phrase dictionary to extract all the medical related phrases. We also remove the high-frequency and low-frequency phrases. In total, we build a dictionary with 2107 phrases.

We use the large corpus of 330,305 threads to build the phrases distribution of the background topic. We also use these 330,305 threads to build a unigram model as a prior for the disease symptom topic of each kind of drug. The disease symptom topic would be further updated by the training collection.

6.2 Qualitative Evaluation

We first show the side effect symptom topics discovered from the health forum using SideEffectPTM. On one hand, some drugs are related to common diseases such as depression so that they appear frequently in the corpus. On the other hand, some drugs appear infrequently in the corpus because they are used to treat some uncommon diseases. Since the online health forum is a mixture of drugs with different frequencies, it is important for our model to
handle all of them. Therefore, we show the results of drugs with high-frequency, middle-frequency and low-frequency. We list each drug with the medical phrases that have the highest $P(w|R_m)$.

From Table 1, we see that SideEffectPTM mines side effect symptom topics successfully for drugs with different frequencies. Most of the mined top symptoms are side effect symptoms instead of the disease symptoms. This proves our model’s ability to distinguish side effect symptoms from disease symptoms by explicitly modeling two parts separately. The extracted symptoms are also general enough to cover a wide range of symptoms from common symptoms, such as headaches and weight gain, to rare but severe symptoms, such as kidney stones and diabetes. Specifically, for the anti-depressant drug Wellbutrin(R), our model extracts seizures as the most likely side effect. Seizures are known as one of the most important ADRs of Wellbutrin(R) in medical literature [8]. Ativan(R) is used to treat anxiety disorders and is reported to cause high blood sugar in 1.16% of patients who have side effects when taking Ativan(R) [9]. Ephedrine only appears in two threads in our data set. However, our model still successfully captures its side effect such as dizziness, tired, tremors, headaches, panic attack and lethargy. These side effects are all verified by the medical literature [11]. The side effect results of Topamax(R) contain several severe symptoms such as liver problems and kidney stones. Recent study shows that treatment with Topamax(R) has been proved to increase the propensity to form calcium phosphate stones [32]. This proves that our model is able to not only mine the common side effects (e.g., headache, weight loss), but also find uncommon severe side effects.

To further demonstrate that our model can separate side effect symptom topics from disease symptom topics, we show both topics of three different drugs in Table 2. From Table 2, we can naturally guess that Seroquel(R) is used to treat depression from the top phrases in its disease symptom topic. On the other hand, in its side effect symptom topic, only the word “anti-psychotic drug” is related to depression. Trazodone(R) is used to treat depression and insomnia. Its disease symptom topic matches this usage with phrases such as “insomnia”, “sleep” and “depression”. Its side effect symptom topic shows its side effect such as “dry mouth” and “high blood sugars”. Celebrex(R) is an anti-inflammatory drug used to treat pain or inflammation. Therefore, its disease symptom topic contains phrases such as “foot”, “neck” and “leg”, which indicate the patients’ pain in different parts of body. However, none of the phrases in its side effect symptom topic is related to pain except its location. Besides these drugs, our model can separate the side effect symptoms from the disease symptoms for most of the drugs and achieve an average cosine similarity of 0.00567. Such a low cosine similarity indicates a large difference between treated topic and side effect topic.

We also extract the most reported side effect symptoms from our experimental results to study the common symptoms. Table 3 is the top 10 most reported side effect symptoms from our experiment results. Most of these symptoms are easy to observe (e.g., weight gain, headaches) or occur frequently in everyone’s daily life (e.g., depression, anxiety). A previous study also discovered this phenomena and explained that people are more conscious of issues that they can directly observe [18]. We think that patients may connect these symptoms with the drug they are using incorrectly because of the anxiety when they are searching for medical information on the internet [33]. At the same time, patients are not able to detect symptoms that can only be observed with the help of physicians or medical devices, such as liver problem or hypertension. These rare unobserved symptoms may be the actual side effect symptoms which cause those common observed symptoms. For example, a liver problem may cause an eating disorder and weight loss. So if a patient reports weight loss and eating disorder, we should also suspect the patient of having a liver problem. Therefore, it is important to mine the potential rare side effects which are hidden behind these common side effects. We propose an approach to mining the rare side effects in Section 5 and apply it to our data set in Section 6.4

### 6.3 Quantitative Evaluation

The main goal of the quantitative evaluation is to see how accurately our model is for mining the side effects. Our proposed model is completely unsupervised which does not require any labeled data.
for training. To perform quantitative evaluation, we construct the following two ground truth sets.

- **Human Annotation**: We first build a human annotation ground truth set. Two Ph.D. students manually extracted the side effects appearing in that thread. Due to the limited manpower, we only labeled threads for four drugs including Zoloft(R), Wellbutrin(R), Tylenol(R) and Topamax(R). These four drugs appear frequently in our collection so that they can provide us more conceivable evaluation results. We labeled 96 threads in total. According to the labeling results, each thread contains 9 side effect symptoms on average.

- **FDA Database**: We use the FDA database as the ground truth set. The FDA database doesn’t always use the same terms as in the health forum, thus there is a mismatch between professional medical term and the plain English term (e.g., dyspnoea) used by patients (e.g., difficulty breathing) [38]. We address this problem by simply doing keyword-matching between mined side effect phrases with those reported side effects to FDA database. Although this is not entirely reliable, it allows us to evaluate our method with many more drug instances. Besides, such a simple matching strategy unlikely introduces any systematic bias, thus it is still useful for performing relative comparisons between different methods.

<table>
<thead>
<tr>
<th>Seroquel(R)</th>
<th>Trazodone(R)</th>
<th>Celebrex(R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sleep</td>
<td>sleep</td>
<td>allergic</td>
</tr>
<tr>
<td>diabetes</td>
<td>depression</td>
<td>surgery</td>
</tr>
<tr>
<td>side effects</td>
<td>anxiety</td>
<td>celebrex</td>
</tr>
<tr>
<td>seroquel</td>
<td>dry mouth</td>
<td>celebrex</td>
</tr>
<tr>
<td>doc prescribed seroquel</td>
<td>tired</td>
<td>breath</td>
</tr>
<tr>
<td>raising blood sugar levels</td>
<td>anti-depressant</td>
<td>drug allergy</td>
</tr>
<tr>
<td>anti-psychotic drug</td>
<td>diagnosed</td>
<td>naprosyn</td>
</tr>
<tr>
<td>ambition</td>
<td>depressed</td>
<td>leg</td>
</tr>
<tr>
<td>mood swings</td>
<td>lithium</td>
<td>short</td>
</tr>
<tr>
<td>diabetic</td>
<td>weight</td>
<td>arthritis</td>
</tr>
<tr>
<td>constipation</td>
<td>weight</td>
<td>knee</td>
</tr>
<tr>
<td>blood work</td>
<td>seroquel</td>
<td>prednisone</td>
</tr>
<tr>
<td>weight</td>
<td>lamictal</td>
<td>pain meds</td>
</tr>
</tbody>
</table>

There is no existing unsupervised approach which mines ADRs from online health forums. We use the following two baselines for comparison.

- **Basic SideEffectPTM (Basic)**: This approach does not incorporate prior in our SideEffectPTM model.

- **PhraseMatch (PM)**: This is the model similar to the approach in [18]. Although they focus on short messages which are less noisy and shorter, we can still apply it to forum text. They assign a score to each phrase in the document according to its similarity to a symptom phrase. We used our dictionary to make this approach match our corpus better. This model is unable to separate side effect symptoms from disease symptoms.

We want to use these two baselines to answer two questions. First, is the prior of disease symptom topic beneficial to our model? We compare Basic SideEffectPTM with our model to answer this question. Second, is it necessary to explicitly build two language models for both side effect symptoms and disease symptoms? We compare our model with PhraseMatch to answer this question.

Since our method would be able to nominate candidate side effects of each drug for human experts to further verify, the task is essentially an ADR retrieval task. Therefore, we can use standard retrieval measures to evaluate our method. We use the following information retrieval measures: Precision, Recall and Mean Average Precision (MAP). From Table 4, we see that our method outperforms all the baselines on all evaluations on both ground truth labeling sets. Our method improves Basic SideEffectPTM by 212% in Precision@3 and 150% in Precision@10 based on the Human Annotation ground truth. This demonstrates that the prior of disease symptom topics can greatly improve the performance when we mine ADRs. Our method improves PhraseMatch by 212% in Precision@3 and 100% in Precision@10. This proves that explicitly modeling side effect symptoms and disease symptoms separately can make the extracted ADRs more accurate.

We also observe that the improvement on FDA database results is less significant than the improvement on human annotation ground truth. Although there are mismatches between the terms in the FDA database and health forum, this may indicate that some of our extracted side effects from the health forum haven’t been approved...
by the FDA. Therefore, we further compare our experiment results with FDA database and find some previously unknown side effects that haven’t been included by FDA in Section 6.5.

6.4 Mining Challenging Adverse Drug Reactions

Besides normal side effects, we also apply our SideEffectPTM model to mining three kinds of more challenging side effects using methods described in Section 5.

From the final top-10 extracted long-term side effect candidates, we detect three potential long-term side effects and show them in Table 5. All of these symptoms have been reported as potential long-term side effects in the medical literatures or in surveys. Phentermine has been reported to cause bone and joint pain after using it for 9 month on average[10]. Below is a patient’s thread about the side effect of Cipro(TM) to him.

Table 5: Long-term Side Effects

<table>
<thead>
<tr>
<th>Drug</th>
<th>Symptom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydroxycoumarin</td>
<td>bleeding</td>
</tr>
<tr>
<td>Cipro(TM)</td>
<td>abdominal pain</td>
</tr>
<tr>
<td>Phentermine</td>
<td>bone and joint pain</td>
</tr>
</tbody>
</table>

Figure 2: Patient’s Thread Reports the Long-term Side Effect of Cipro(TM)

Table 6 shows the top extracted drug-drug interactions we have mined through our model. Most of these drug pairs have been reported to be dangerous when taken together. Methadone and Xanax(R) are reported to have potentially life-threatening effects if taken together[15]. A recent review of Methamphetamine suggests that Zoloft(R) might not be effective in the treatment of methamphetamine addiction and might even worsen the condition[17]. Valium(TM) has been reported as one of the major interactions with Zoloft(R) and the level of medication in the bloodstream should be monitored closely when taken together[25].

Table 7 is the extracted rare side effects. These top extracted drug-symptom pairs reveal some potential rare side effects. For example, Prozac(R) has been reported to have the sexual side effect, especially ejaculation problems[7]. The anti-inflammatory drug Solaraze(R) has been reported to cause rare but significant cases of serious hepatotoxicity in medical literature[5].

To further demonstrate the effectiveness of our model and prove that health forums are good information sources to collect ADRs, we compare the top extracted ADRs with the FDA database. Since there is a mismatch between phrases in two corpus (e.g., “difficulty breathing” in forum, “dyspnoea” in FDA database) [38], we manually check the extracted ADRs. Table 8 is the potential ADRs listed in our experiment results but not listed in the FDA database.

Table 8: Discovering Unreported ADRs

<table>
<thead>
<tr>
<th>Drug</th>
<th>Symptom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prozac(R)</td>
<td>ejaculate</td>
</tr>
<tr>
<td>Ativan(R)</td>
<td>raising blood sugar levels, diabetic, constipation</td>
</tr>
<tr>
<td>Adderall(TM)</td>
<td>lost weight</td>
</tr>
</tbody>
</table>

We notice that the FDA database often covers more ADRs than the ADRs mined from health forums by using our proposed model. However, our approach still has the following advantages. First, our model could successfully mine some previously unknown ADRs that haven’t been collected by FDA database. Second, although we return fewer ADRs, we still maintain results of high quality and accuracy based on both qualitative and quantitative evaluations. In addition, our model can mine ADRs for new drugs more efficiently than the FDA database and link the ADRs to the corresponding documents.

7. CONCLUSION AND FUTURE WORK

In this paper, we use the abundant text information from online health forum to mine adverse drug reactions. Our proposed SideEffectPTM is the first approach that separates side effect symptoms from disease symptoms when mining ADRs. We successfully
mine many meaningful ADRs from real-world online health forum. Some of them have never been reported in FDA database. Furthermore, we also mine some more challenging side effects such as long-term side effects, rare side effects and drug-drug interactions, which are generally very hard to detect.

The voluntary participation of huge amounts of patients and the rapid growth of online health forums will bring us more and more forum text data. Our method can be continuously applied to these growing data to discover new ADRs reported and link the extracted ADRs to the corresponding documents. Our approach can be further improved by addressing the data quality problem of the forums, such as misspelled words, different forms of the same word and unrelated information being brought up in a single thread.

8. REFERENCES