Journal of Biomedical Informatics 54 (2015) 230–240

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Filtering big data from social media – Building an early warning system for adverse drug reactions

Ming Yang [a,*], Melody Kiang [b], Wei Shang [c]

[a] Department of Information Management, School of Information, Central University of Finance and Economics, Beijing 100081, China
[b] Department of Information Systems, California State University, Long Beach, CA 90840, United States
[c] Academy of Mathematics and Systems Science, Beijing 100190, China

ARTICLE INFO

Article history:
Received 8 August 2014
Accepted 27 January 2015
Available online 14 February 2015

Keywords:
Partially supervised classification
Latent Dirichlet Allocation (LDA)
Adverse drug reactions
Social media filtering
Social media mining

ABSTRACT

Objectives: Adverse drug reactions (ADRs) are believed to be a leading cause of death in the world. Pharmacovigilance systems are aimed at early detection of ADRs. With the popularity of social media, Web forums and discussion boards become important sources of data for consumers to share their drug use experience, as a result may provide useful information on drugs and their adverse reactions. In this study, we propose an automated ADR related posts filtering mechanism using text classification methods. In real-life settings, ADR related messages are highly distributed in social media, while non-ADR related messages are unspecific and topically diverse. It is expensive to manually label a large amount of ADR related messages (positive examples) and non-ADR related messages (negative examples) to train classification systems. To mitigate this challenge, we examine the use of a partially supervised learning classification method to automate the process.

Methods: We propose a novel pharmacovigilance system leveraging a Latent Dirichlet Allocation modeling module and a partially supervised classification approach. We select drugs with more than 500 threads of discussion, and collect all the original posts and comments of these drugs using an automatic Web spidering program as the text corpus. Various classifiers were trained by varying the number of positive examples and the number of topics. The trained classifiers were applied to 3000 posts published over 60 days. Top-ranked posts from each classifier were pooled and the resulting set of 300 posts was reviewed by a domain expert to evaluate the classifiers.

Results: Compare to the alternative approaches using supervised learning methods and three general purpose partially supervised learning methods, our approach performs significantly better in terms of precision, recall, and the F measure (the harmonic mean of precision and recall), based on a computational experiment using online discussion threads from Medhelp.

Conclusions: Our design provides satisfactory performance in identifying ADR related posts for post-marketing drug surveillance. The overall design of our system also points out a potentially fruitful direction for building other early warning systems that need to filter big data from social media networks.

1. Introduction

The safety of medicines is a major concern for patients. Harmful, unintended reactions to medicines that occur at doses normally used for treatment are called adverse drug reactions (ADRs). ADRs are among the leading causes of death in many countries. Since 1960s ADRs have been monitored in many countries and by the World Health Organization (WHO) using pharmacovigilance systems, also called "early warning" systems [1]. The primary aim of these systems is to collect information about possible ADRs, particularly for serious, rare, and unknown ADRs, at an early stage after the drugs were marketed. During the clinical trials, that are usually carried out in the evaluation and marketing authorization stages, the safety of drugs can only be investigated to a limited extent. Therefore, it is essential to monitor the safety of drugs after marketing [2].

Typically, pharmacovigilance systems rely on the reporting by physicians and pharmacists, not directly from the patients. Therefore, the reports that reach the pharmacovgilance system may not reflect the adverse events that were originally reported because of the filtering effect of physicians and pharmacists. With the increase of patients' understanding of illness, many patients wish

* Corresponding author.
E-mail addresses: yangming@cufe.edu.cn (M. Yang), mkiang@csulb.edu (M. Kiang), shangwei@amss.ac.cn (W. Shang).

http://dx.doi.org/10.1016/j.jbi.2015.01.011
1532-0464/

to be more involved in decisions regarding his or her disease and therapy. Pharmaceutical companies are also interested in the direct reporting of ADRs by consumers in a timely manner during post-marketing drug surveillance due to the severe legal and monetary implications [2]. Since reporting of ADRs by patients is in line with the striving for quality in the healthcare system, a growing number of countries allow patients to report suspicious ADRs directly to a pharmacovigilance system [3]. Study has shown that consumer reporting of ADRs contributes significantly to a reliable pharmacovigilance [4]. However, not all countries accept consumer reports, especially for developing countries where around 80% of the global population resides [5]. Also a considerable time lag exists in recognition of serious ADRs using the consumer reporting. Hence, there is a need for a different approach to the existing pharmacovigilance.

Social media provides patients a platform to exchange their drug use experiences. Moreover, social media constitutes a significant part of the online search results for information about health and medical matters [6]. Healthcare research could benefit from taking advantage of this rich information resource [7,8]. Van Hunsel et al. [9] investigated the motives for reporting ADRs by patients in the Netherlands, showing that patients are willing to share their experiences regarding the use of drugs on social media. These user-generated content (UGC) is rapidly emerging as tremendous assets for syndromic surveillance, which is concerned with the continuous monitoring of public health-related sources and early detection of adverse disease events [10]. Moreover, previous research has shown that the analysis of patients' narratives posted on social media websites is important for assessing the consumers' perceived risk of ADRs [11], and mining the relationship between drugs and adverse reactions [12,13]. Finding and analyzing consumer-generated ADR messages, buried among millions of consumer posts, is a challenge that has received very limited attention in prior literature. How to effectively gathering the vast amounts of drug use information generated by consumers, and sifting out the ADRs related messages, is the focus of this research.

However, filtering the consumer ADRs related messages from social media is not a trivial task. The challenge is: consumer ADR related messages are usually sparse and highly distributed, while non-ADR messages are unspecific and topically diverse. It is costly and time consuming to manually classify and label a large number of consumer ADR messages and non-ADR messages for building early warning systems. Nevertheless, it is relatively easy to obtain large amounts of unlabeled content on social media. Our research endeavors to develop a new process to scan large amount of text-based posts collected from drug-related Web forums. The proposed system integrates both text and data mining techniques to automatically extract important text features from the posts first, and then classify the posts into positive/negative examples based on a few pre-identified ADR related posts. The classification process is based on a partially supervised learning method, which uses a small number of known positive posts to identify other posts of similar text features from a large corpus of unlabeled posts. We test our method on drug-related Web forums and the preliminary results are encouraging. The proposed method can assist Food & Drug Administration (FDA) and pharmaceutical companies in identifying suspicious ADR messages on social media and the result can be used as input to build an early warning system to prevent future ADRs.

The remainder of the paper is organized as follows. Section 2 provides the background for text mining techniques in syndromic surveillance and existing partially supervised learning methods. We also summarize current research gaps and the need of our study. Sections 3 and 4 present the experimental methods and the discussion of the results. Section 5 concludes our discussion with a summary of our contributions and suggestions for future research directions.

2. Research background

2.1. Text mining in syndromic surveillance

Text mining techniques have been widely deployed for text classification in a wide spectrum of public healthcare problems. For example, Lu et al. [14] proposed an ontology-enhanced approach for classifying free-text chief complaints (CCs) from the emergency department. Botsis et al. [15] employed a multi-level text mining approach for automated text classification of VAERS (Vaccine Adverse Event Reporting System) reports. In order to detect early indications of disease outbreaks from online news, researchers employed text classification in Internet-based bio-surveillance projects [16,17].

A significant amount of research has been done in trying to identify high-quality healthcare information in social media [18]. For instance, Denecke and Nejdl [19] compared the content of medical Question & Answer Portals, medical Weblogs, medical reviews, and Wikis. The results showed that there are substantial differences in the content of those health related social media. Huh et al. [20] applied text classification methods to determine whether a thread in an online health forum needs moderators' help. Based on the use of tags and tag clouds, O'Grady et al. [21] assessed the credibility of messages in online health forums. Chee et al. [22] used a machine learning method to classify drugs into FDA's watch list and non-watch list based on messages extracted from online health forum.

The fundamental approach in previous studies for syndromic surveillance using text based data source was mostly information retrieval, including ad hoc retrieval and text categorization. Ad hoc retrieval refers to retrieving text from a relatively static text collection in response to short term queries. Text categories are predefined according to the long-term information needs of the users. For those studies, examples of documents labeled with preference categories are often available, therefore the task is usually casted as a supervised classification problem [23].

2.2. Partially supervised classification

Supervised learning algorithms require high-quality labeled training data in order to construct an accurate classifier. However, messages related to consumer ADRs are usually topically diverse and highly distributed in social media. It is often a mentally exhausting, if not infeasible, process to manually acquire and label a large number of consumer ADR posts in order to train a classifier. In addition, reliable and up-to-date health-related data are of varying quality, and are difficult to locate on the Web [24]. Finally, due to the dynamically changing environment of social media, the labeled training data may soon become outdated.

One way to overcome the difficulties is to dynamically augment the training data through a partially supervised learning algorithm, which constructs classifiers based on mostly unlabeled data and a small number of labeled positive examples that are of interest to the users. Fung et al. [25] summarized the characteristics of partially supervised learning as follows: (a) the size of the given positive examples is so small that it might not be possible to represent the feature distribution of all positive examples, (b) the unlabeled examples are mixed with both positive and negative examples, and (c) no negative example is given. Since no negative example is given explicitly, it is critical to design good labeling heuristics (i.e., models/features/kernels/similarity functions) for identifying both positive and negative examples from the unlabeled datasets [26]. Generally speaking, the existing approaches that target this

problem can be categorized into three classes: (1) enlarging reliable negative training examples, (2) enlarging reliable positive training examples, and (3) enlarging both positive and negative training examples.

The enlarging reliable negative training examples method tries to extract representative negative examples from the unlabeled dataset and train the classifier based on the given positive examples and the extracted negative examples. Yu et al. [27] proposed an approach called PEBL (Positive Example Based Learning) that utilizes Support Vector Machine (SVM) to construct a classification model. PEBL first employs a rough classifier to initialize approximation of "strong negative" examples, and then constructs an initial classifier using the "strong negative" and positive examples. Next, PEBL iteratively uses the obtained model to identify more negative examples so as to induce a classification model capable of differentiating the boundary between positive and negative classes. Li and Liu [28] proposed a technique based on the Rocchio algorithm and SVM. This approach utilizes the Rocchio algorithm for classification and clustering to extract a set of reliable negative documents from the unlabeled set, and then constructs a classifier iteratively using SVM. Based on the assumption that the acquisition of the legitimate emails (negative examples) is more difficult than that of spam (positive examples), Wei et al. [29] proposed an E2 technique following the PEBL framework for spam filtering. E2 follows the two-stage framework of PEBL but extends each stage with an ensemble strategy. The effectiveness of these methods relies on the premise that the labeled positive examples are sufficient to capture the diverse characteristics of the positive class.

In many information retrieval applications, positive examples refer to the data points that are of interest to the researchers in a binary classification problem. For instance, in spam detection, researchers often employ spam as positive examples and utilize legitimate emails as negative examples. Moreover, the boundary (definition) of a positive class is usually more specific than that of a negative class [30]. Hence, it is more likely to identify potential positive examples from the unlabeled dataset through exploiting the inherent structures in the positive examples. Ko and Lam [31] proposed a technique called EAT (Example Adaption for Text categorization) for automatically seeking more representative positive examples from the unlabeled documents. This approach consists of two steps: first, extracting a set of potentially positive examples from an unlabeled dataset; second, generating a set of classifiers iteratively through gradually increasing the number of positive examples until the classifier reaches its local maximum accuracy level. The effectiveness of EAT is based on the content-specific features that capture the characteristics of the positive class. However, the content-specific features of EAT are manually crafted from a very small number of sample documents. Thus, the effectiveness of the classifier is highly dependent on the quality of the content-specific features given.

Nevertheless, it is not an easy task to extract a proper set of positive examples due to the diversity of topics exhibited in unlabeled messages. In order to solve this problem, Fung et al. [25] proposed an approach called PNLH (Positive examples and Negative examples Labeling Heuristics) using partition-based heuristics that iteratively extract reliable positive and negative examples from an unlabeled dataset. The effectiveness of this approach depends on the core vocabularies of the positive examples (i.e., positive features). To be more specific, the underlying assumption of this approach is that the positive features are sufficient to capture the characteristics of the positive class. When the initial labeled training dataset is too small to be representative of the true positive class, the performance of this method may deteriorate. In order to solve this problem, Zhang and Xiao [32] developed an algorithm called ACTC (Active semi-supervised Clustering based Two-stage text Classification). Using the labeled data as the guide, this approach first clusters both labeled and unlabeled data. During the process, a self-training style clustering strategy is used to iteratively expand the training. At the second stage, a discriminative classifier is trained with the expanded labeled dataset.

Although consumer ADR messages are more specific than non-ADR messages, they are still diverse in topics, including different drugs, side effects, and diseases. As every topic contains its own set of core vocabulary, a large number of different topics cancels out the significance of each others' core vocabulary [25]. Eventually, it is difficult to extract reliable positive and negative examples based on the core vocabulary. Moreover, in the problem of partially supervised classification, the most commonly used feature space is the term space which is usually of very high dimension. The performance of the partially supervised classification may deteriorate due to the high-dimensionality of the text data. However, previous studies have placed limited emphasis on the dimension reduction in partially supervised classification.

2.3. Research opportunities and objectives

Our review of text mining techniques in syndromic surveillance and existing partially supervised learning methods reveals several research opportunities. First, limited research has investigated the role of partially supervised classification in early warning systems for ADR. Second, existing partially supervised classification research provides limited support for filtering the consumer ADRs related messages from social media.

Based on these observations, our research is aimed at: (a) developing an early warning system which leverages existing partially supervised classification research and (b) gaining an understanding of the importance of consumer ADRs related messages from social media in post-marketing surveillance. The objective of our research is to bridge the technical gaps existing in the current ADR detection and to develop an efficient automatic syndromic classification method that can filter consumer ADRs related messages from social media.

3. Experimental methods

This study is aimed at designing and examining a new approach to identify consumer ADR posts on social media. Specifically, we develop and evaluate informatics tools and frameworks using a partially supervised classification approach to monitor content that contains negative sentiments toward certain drugs at targeted Web forums. Such sentiments could be important indicators of potential adverse drug reactions [22]. Fig. 1 shows the overview of the proposed framework. The proposed framework allows for an automated filtering of postings, where the text classification system is built on LDA modeling and augmenting training data using partially supervised learning. The detailed description of each process is presented in the following sub-sections.

3.1. Data collection

We begin with the crawling of Web forums to gather patient posts from social media. In this study, we confine the analysis to discussion boards only; however, the same process can be extended to other social media sources like Twitter and Facebook. Parsing programs are then developed to extract important information from the raw Web pages and store them in a relational database. We extract the following information from each thread: title, text of each post, date and time of each post, user identifier of each post, and number of times the thread has been viewed.

Based on the annotation guidelines in [33], we recruited domain experts to tag a large sample of posts related to the drugs in our study. The experts read each post assigned to them and assess

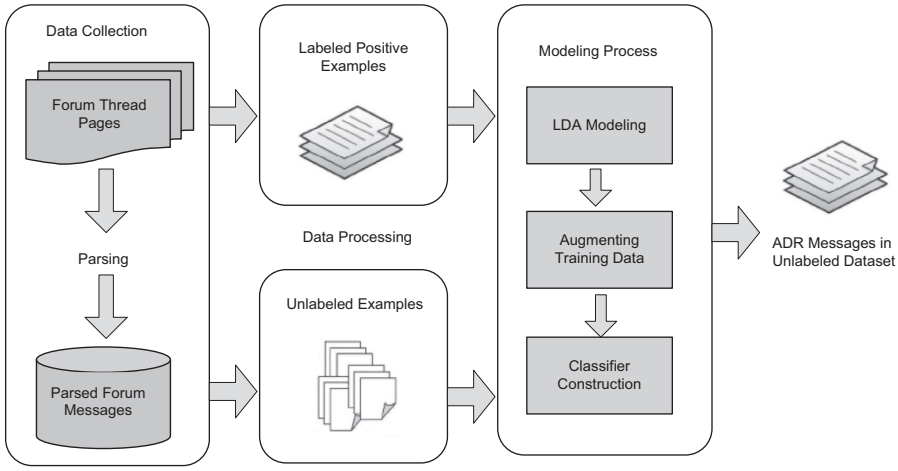

Fig. 1. Overview of the early warning system design based on social media data.

whether any ADR has been discussed. If a post contains any ADR related wordings, it is labeled as a positive example; otherwise, it is labeled as a negative example.

Specifically, we collect online discussion threads of drugs from Medhelp. As the pioneer in online health community, Medhelp has over 12 million users each month sharing medical information and finding answers to their medical questions since its inception in February 1994 [13]. In the Drugs section of MedHelp, users can start a thread of certain drugs with a post, on which other users can comment. There are tens of thousands of drugs in the Drugs section,[1] and there can be thousands of threads under each drug. We follow a previous research in ADR detection [13], and use the keywords (e.g., Biaxin, Lansoprazole, and Luvox) to spider 3500 threads of discussion from Medhelp between January 2004 and April 2014. Three independent medical domain experts are employed to screen and select ADR posts and non-ADR posts. First, the training set is tagged by two of the domain experts. To ensure unbiased result, each expert is asked to independently tag 1500 posts (500 Biaxin posts, 500 Lansorprazole posts, and 500 Luvox posts). Tagging inconsistencies are then noted and the experts construct a protocol document to govern the tagging process. Next, the experts adjust their discrepant tags to conform to the agreed protocol. After all training posts are tagged, a Kappa statistic of inter-rater reliability is computed. For the "ADR vs. non-ADR" variable, the Kappa measure is 0.89, which indicates a very strong agreement between the two domain experts and the outcome is reliable. Finally, the two experts are asked to resolve their remaining differences to construct a gold standard training set for the subsequent analysis. The testing set is then independently tagged following the same protocol by the third domain expert.

Preprocessing is performed on all three datasets. First, punctuation, numbers, non-alphabet characters, and stop words are removed. Second, standard stemming is performed to reduce inflected or derived words to their stem, thus reduce the vocabulary size and address the issue of data sparseness. Descriptive statistics of the datasets before and after preprocessing are shown in Table 1.

Table 1
Descriptive statistics of the datasets.

Dataset	# Of words		
	Biaxin	Lansoprazole	Luvox
Average doc. length†	184	171	179
Average doc. length*	116	113	108
Vocabulary size†	23,941	21,534	20,138
Vocabulary size*	19,893	20,867	18,643

† Denotes before preprocessing.
* Denotes after preprocessing.

3.2. Feature space modeling

A feature space is an abstract space where each data instance is represented as a point in an n-dimensional space. Its dimension is determined by the number of features used to describe the instances. In order to characterize the diversity of the consumer ADR posts, we need to utilize a large corpus in order to achieve better coverage. Since the consumer corpus contains hundreds of thousands of terms, the challenge is to reduce the high dimensionality of the feature space while maintaining the semantic structure of the original vector space. To this end, we apply the Latent Dirichlet Allocation (LDA) [34] to construct a topic space over the corpus. The LDA is a generative probabilistic model that uses a small number of topics to describe a collection of documents. By representing a document in the topic space instead of in the term space, the LDA model can effectively reduce the dimension of the texts while maintaining the semantic structure of the document.

In LDA, each document in the corpus is modeled as a set of draws from a mixture distribution over a set of hidden topics. A topic is modeled as a probability distribution over words. Let T be the number of topics a LDA model and V be the size of the vocabulary of the corpus. The LDA model simulates the generation of a document with the following stochastic process: (1) For a document $d_m = \{w_{m,n}\}_{n=1}^{N}$, sample a topic proportion vector $\theta = (\theta_1, \theta_2, \ldots \theta_T)'$ from a Dirichlet distribution Dir(α). This is equivalent to an author deciding what topics to include in a paper. (2) For a word $w_{m,n}$ in the document d_m, sample a topic $z_{m,n}$ according to the multinomial distribution $Mult(\theta_m)$. This process can be regarded as assigning a word to a topic. The parameter θ_m is a V-dimensional vector that defines the multinomial word-usage distribution of a topic. It is distributed as Dirichlet with parameter β. (3) Conditioning on a topic $z_{m,n}$, sample a word $w_{m,n}$ according to the multinomial distribution $Mult(\phi_{z_{m,n}})$. This corresponds to picking words to represent a concept. In order to estimate the parameters θ and ϕ for LDA, we need to maximize the likelihood of the whole data collection – i.e., the entire corpus D:

$$p(D|\alpha,\beta) = \prod_{d_m \in D} p(d_m|\alpha,\beta) = \iint p(\phi|\beta) \prod_{n=1}^{N} p(w_{m,n}|\phi_{z_{m,n}}) p(z_{m,n}|\theta_m) p(\theta_m|\alpha) d\phi d\theta_m \quad (1)$$

[1] http://www.medhelp.org/health_topics/drugs_list.

However, the two integrations in Eq. (1) are intractable. We use the Gibbs sampling inference algorithm [35] for approximation here, because it is less likely to be trapped in a local maxima than the other approaches. Let z denotes the vector of the instances of all latent topic variables and w denotes the vector of all the observed words of the corpus. The inference algorithm concentrates on the joint probability $p(w,z)$ and applies a sampling approach to instantiate the latent topic variable for each word. Gibbs sampling is a technique to generate samples from a complex posterior distribution $p(z|w)$ by iteratively sampling and updating each latent variable z_i according to the conditional distribution $p(z_i|w,z_{-i})$, where z_{-i} denotes the current instantiation of all the latent topic variables except z_i and w denotes the vector of all observed words of the corpus. The Gibbs sampling algorithm is as follows: (1) the latent variables z are first randomly initialized; (2) each element z_i of z is iteratively sampled and updated; (3) repeat the last step until the Markov chain converges to the target posterior distribution $p(z|w)$; (4) samples of z can be collected from the Markov chain.

3.3. Augmenting training data using partially supervised learning

In this research, we aim to identify consumer posts (documents) on health related Web forums as candidates for the watchlist (ADR related) with relatively few known positive (watchlist) examples. Specifically, we try to use a small subset of positive labeled posts to find other posts of the same class within a large corpus of unlabeled posts. A small number of positive examples tend to lead a classifier to over-fitting. However, manually annotating a large amount of training examples is an expensive process and does not scale well for the large dataset in our case. To overcome the difficulty associated with scarce positive training data, the proposed method can automatically augment the training data using partially supervised learning that extracts candidate positive and negative examples from the unlabeled dataset. Hence, it is critical to design a labeling heuristic which is sufficiently smooth with respect to the intrinsic structure revealed by the labeled and unlabeled points.

In feature space modeling, we use LDA to summarize the posts $\{d_m\}_{m=1}^{M}$ in the whole corpus to a few latent topics $\{z_j\}_{j=1}^{T}$. When learning parameters θ and ϕ in Eq. (1), we obtain the topic association of each word in the posts in our corpus. Thus, we are able to obtain the post-topic association and the word-topic association, represented as $p(z_j|d_m)$ and $p(w|z_j)$, respectively. We can also obtain the conditional probability of generating post d_m from a topic z_j using the Bayes rule:

$$p(d_m|z_j) = \frac{p(z_j|d_m)p(d_m)}{\sum_{m=1}^{N} p(z_j|d_m)p(d_m)} \tag{2}$$

where $p(z_j|d_m)$ is the topic proportion provided by the topic model. We assume a uniform prior probability distribution for $p(d_m)$.

Using the word-topic association $p(w|z_j)$, we can also infer the hidden topic for the incoming new post x:

$$p(z_j|x) = \frac{p(x|z_j)p(z_j)}{\sum_{j=1}^{T} p(d'|z_j)p(z_j)} = \frac{p(z_j)\prod_{w\in d} p(w|z_j)}{Z} \tag{3}$$

where $p(z_j)$ is the prior probability of the hidden topic z_j, and Z is the normalization factor. Then we are able to obtain the similarity between a new post x and the post d_m in the corpus:

$$p(d_m|x) = \sum_{j=1}^{T} p(d_m|z_j)p(z_j|x) \tag{4}$$

Consider an example to illustrate how to calculate the similarity between the post d_m in the corpus and an incoming post x. We assume that the latent topics might refer to the four aspects: <"Diarrhea", "Heart Disease", "Depression", "adverse drug reaction">. Consider one post with the title "Not sure what's wrong", which describes the ADR after taking Biaxin. During the learning process of the topic model, each word in this post is assigned to one of the four latent topics. After normalization, the topic distribution of this document can be represented by a vector <0.4, 0.3, 0.3, 0.7>, which denotes the strength of the soft association between this post and the four latent topics. Similarly, when another post talking about "Taking Biaxin for Years" comes, using the same word-topic association (3), it can be represented with the latent topic vectors as <0.6, 0.2, 0.2, 0.7>. Finally, we can calculate the similarity between "Not sure what's wrong" and "Taking Biaxin for Years" on the latent topic space using (4), that is $0.4*0.6+0.3*0.2+0.3*0.2+0.7*0.7=0.85$.

In order to evaluate the relatedness of a post with the labeled positive examples, we aggregate the conditional distribution of all data in the same domain D^L:

$$p(d_m|D^L) = \frac{1}{Z}\sum_{x_j^L\in D^L} p(d_m|x_j^L) \tag{5}$$

where D^L is the labeled positive example set. We form the candidate positive dataset by extracting the top M ranked candidates which are related to D^L according to Eq. (5). We use the conditional probability $p(d_m|D^L)$ as the labeling confidence, estimating how close an unlabeled example is to the labeled positive dataset. Fig. 2 shows how to extract more candidate positive examples and negative examples from the unlabeled dataset.

3.4. Classifier construction

The extracted candidate positive and negative examples are not always reliable. When too many candidate examples are extracted, it may cause over-fitting problem and lead to performance degradation [36]. Hence, during the classifier construction, we select reliable positive and negative examples to fit the distribution of the positive and negative classes, respectively.

Recall that candidate negative dataset C_N consists of diverse topics. It is such diversity that makes all approaches that we discussed so far impractical to extract reliable examples from the unlabeled set. The key issue is therefore to reduce the diversity of topics. One way to approach the problem is to partition C_N into smaller clusters, each represents a fewer number of topics. In the following, we discuss how to solve this problem through two separate procedures, namely, purifying candidate negative dataset, and building classifier by iteratively selecting reliable positive examples.

To purify C_N, we randomly divide C_N into smaller clusters. Using each cluster and the labeled positive set D^L as input, we build classifier to identify reliable negative examples in the cluster and extract them. The idea is to identify reliable negative examples from C_N in a localized manner. By doing so, each partition focuses on a small set of more related features.

Many existing clustering algorithms are available for partitioning C_N. In this study we adopt the classical k-means clustering algorithm due to the popularity and the simplicity of the method. A common question when applying k-means clustering is regarding how to determine the optimal number of clusters, k. The results from our preliminary study (see Section 4.2) shows the performance of the proposed system is not sensitive to the selection of k as long as k is not too small. When there is not enough number of clusters to effectively represent different groups of topics in the post, the within cluster diverse topic problem still exist. The detailed steps of the purifying procedure are described in Fig. 3.

Input: Labeled positive dataset D^L, unlabeled dataset U, and relevance threshold δ.

Output: Candidate positive dataset C_P, candidate negative dataset C_N.

1. Initialize: Set $C_p = \varnothing$, $C_N = U$.
2. **for** each d_m in U **do**
3. **if** $p(d_m \mid D^L) \geq \delta$ **then**
4. Let $C_p = C_p \cup \{d_m\}$, $C_N = U - \{d_m\}$.
5. **end if**
6. **end for**

Fig. 2. Algorithm for selecting candidate positive and negative examples.

Input: Candidate negative dataset C_N, Labeled positive dataset D^L.

Output: Reliable negative dataset R_N.

1. Initialize: Set $R_N = \varnothing$.
2. Choose k initial cluster centers $\{O_1, O_2, ..., O_k\}$ randomly from C_N.
3. Perform k-means clustering to produce k clusters, i.e., $C_{N1}, C_{N2}, ..., C_{Nk}$.
4. **for** t = 1 to k **do**

   Two prototype topic vectors, $\vec{P}_t$ and $\vec{N}_t$, corresponding to the positive and the negative prototype respectively, are learned by the Rocchio algorithm as follows:

$$\vec{P}_t = \alpha' \frac{1}{|D^L|} \sum_{d_m \in D^L} \frac{\vec{d}_m}{\|\vec{d}_m\|} - \beta' \frac{1}{|C_{Nt}|} \sum_{d_m \in C_{Nt}} \frac{\vec{d}_m}{\|\vec{d}_m\|}$$

$$\vec{N}_t = \alpha' \frac{1}{|C_{Nt}|} \sum_{d_m \in C_{Nt}} \frac{\vec{d}_m}{\|\vec{d}_m\|} - \beta' \frac{1}{|D^L|} \sum_{d_m \in D^L} \frac{\vec{d}_m}{\|\vec{d}_m\|}$$

5. **for** each d_m in C_N **do**

   find the nearest positive prototype topic vector $\vec{P}_t$ to $\vec{d}_m$

6. **if** there exist a $\vec{N}_t$ (t=1,2,…k), *s.t.*

   $\vec{d}_m \bullet \vec{P}_t \leq \vec{d}_m \bullet \vec{N}_t$ **then**

   $R_N = R_N \cup \{d_m\}$

7. **end if**
8. **end for**

Fig. 3. Procedure to identify reliable negative examples.

After extracting all reliable negative dataset R_N from U, we begin to select reliable positive examples from C_P. The whole selection process is outlined in Fig. 4. We select reliable positive examples by iteratively deleting noisy data from C_P. We obtain a set of reliable positive dataset R_P by merging the given D^L and the updated positive dataset R'_P. The final classifier is constructed by running a particular classifier iteratively with different sizes of R_P. The classifier C_m is selected based on its local maximum F_1 score for positive class on the dynamically updated training data. We choose Support Vector Machines (SVM) as the text classifier due to its popularity and superb performance in text classification.

4. Empirical evaluation

In this section, we demonstrate the effectiveness of the proposed approach for consumer ADR post identification. We first describe our evaluation framework, which includes the test beds and the evaluation criteria. This is followed by an analysis of the results.

4.1. Evaluation framework

For each dataset, we randomly select 40% of the ADR posts as positive training examples, and mix another randomly selected 40% of the ADR posts and 40% of the non-ADR posts as unlabeled examples. To maintain the same class distribution, we create our test set from the remaining 20% of the ADR posts and the randomly selected 20% of non-ADR posts. Table 2 shows the composition of the three test beds.

The performance is compared using the F_1 score on positive examples (ADR posts). The F_1 score is a popular measure used in information retrieval and machine learning which trades off precision p and recall r. The precision p and recall r are defined as follows:

$$p = \frac{\text{Number of true positive posts}}{\text{Number of the posts classified as positive}} \quad (6)$$

$$r = \frac{\text{Number of true positive posts}}{\text{Total number of positive posts in fact}} \quad (7)$$

Input: Labeled positive dataset D^L, unlabeled dataset U, and relevance threshold δ.

Output: Reliable positive dataset R'_P.

1. **Preprocess**: obtain candidate positive dataset C_P and candidate negative dataset C_N via Algorithm 1.
2. Train a classifier C_m using C_P as the positive training set and C_N as the negative training set, F_0 is the F-measure of C_0 for the positive class.
3. **Initialize** $F_{\max} = F_0; m = 1; R'_P = \varnothing$.
4. Obtain R_N through running Algorithm 2.
5. **Repeat**
6. **if** $C_P \neq \varnothing$ **then**
7. Let $\delta = \delta + \Delta\delta$.
8. **for** each d_m in C_P **do**
9. **if** $p(d_m \mid D^L) \leq \delta$ **then**
10. Let $C_P = C_P - \{d_m\}, R'_P = C_P$.
11. **end if**
12. **end for**
13. **end if**
14. Let $R_P = D^L \cup R'_P$.
15. Construct classifier C_m using R_P as the positive training set and R_N as the negative training set.
16. Let F_m be the F-measure of C_m for the positive class.
17. **if** $F_m \geq F_{\max}$ **then**
18. Let $F_{\max} = F_m$.
19. **end if**
20. m++;
21. **until** $F_{m-1} < F_{\max}$.
22. **return** R'_P.

Fig. 4. Algorithm for selecting reliable positive examples.

F_1 score is the harmonic mean of precision p and recall r:

$$F_1 = \frac{2pr}{p+r} \tag{8}$$

To obtain a high F_1 score, both precision and recall need to be high.

GibbsLDA++ package is applied to project the domain data onto the latent topic space, using the posts in the three test beds collected. In the LDA model implementation, we set the symmetric prior $\beta = 0.01$ as suggested in [37]. The asymmetric prior α is learned directly from data using maximum-likelihood estimation [38], and updated every 25 iterations during the Gibbs sampling procedure.

4.2. Parameter tuning

We first conduct preliminary runs to determine appropriate values for the three key parameters: T (the number of topics), k (the number of clusters), and $\Delta\delta$ (the increment of the relevance threshold). We use the training examples as the tuning set to determine suitable parameter values.

In this research, we apply the LDA model to construct a topic space over the corpus, and simultaneously train the partially supervised classification approach on the low-dimensional representations generated by LDA. The core assumption of using the LDA model is that the dimensionality of the topic variable z needs to be known prior to the training. It is therefore worth exploring how the change in the number of topics influences the performance of the proposed method for ADR post identification. With

Table 2
Composition of the three test beds.

Dataset	# Of posts (documents)		
	Biaxin	Lansoprazole	Luvox
Corpus			
ADR posts	1000	1000	1000
Non-ADR posts	4000	4000	4000
Training set			
Labeled positive examples[a]	400	400	400
Unlabeled examples[b]	2000	2000	2000
Testing set			
Labeled positive examples[c]	30–80	30–80	30–80
Unlabeled examples[d]	970–920	970–920	970–920

[a] 40% of the ADR posts.
[b] 40% of the ADR posts and 40% of the non-ADR posts.
[c] 15–40% of the remaining 20% ADR posts.
[d] 85–60% of the remaining 20% ADR posts and 20% of the remaining non-ADR posts.

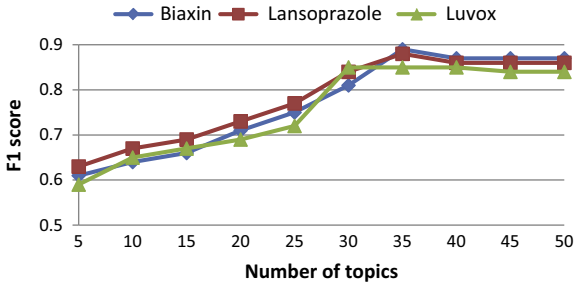

Fig. 5. ADR post identification accuracy using different number of topics.

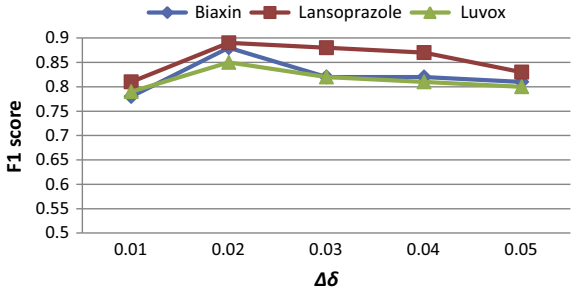

Fig. 7. The effect of parameter $\Delta\delta$ on F1 score of the extracted ADR posts.

this in mind, we conduct a set of experiments, with different topic numbers $T \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. Fig. 5 shows the ADR post identification results with different number of topics for the three drugs selected for studying. As can be seen in Fig. 5, the performance of the proposed method improves as the number of topics (T) increases when $T < 35$. When the number of topics is set to 35, the proposed method performs the best on all three test beds. The performance of the proposed method becomes less sensitive to the change in the number of topics when $T > 35$. The experimental results show that by applying the LDA model to reduce the dimensionality problem, the resulting low-dimensional semantic representation can still effectively characterize the diversity of the consumer ADR posts.

The non-ADR posts (negative examples) comprise more diverse topics than the ADR posts. We use the k-means clustering algorithm to cluster and purify the candidate negative dataset, which allows us to identify and remove noisy data in C_N in a localized manner. It is important to select an appropriate number of clusters to capture the diversity of the negative class. Hence, we perform test runs to tune parameter k to find the best number of clusters. Fig. 6 shows the ADR post identification results using different number of clusters. The number of clusters k varies from 2 to 30. We observe that for all three test beds, the performance of the proposed method is less sensitive to the choice of k value when k is greater than 10.

To better divide the ADR posts and non-ADR posts, we construct the final classifier by running a particular categorization scheme iteratively using different sizes of R_P. This is to ensure that a significant number of reliable positive examples are extracted from the unlabeled dataset. To facilitate the classifier construction process, we introduce parameter $\Delta\delta$ to control the change of the relevance threshold in the iterative process that converges. Smaller $\Delta\delta$ value will make the change of the relevance threshold slower, and as a result the algorithm will take longer to converge. Larger $\Delta\delta$ value will make the relevance threshold decay faster, and might result in missing the optimum value. We set the value of $\Delta\delta$ to a range between 0.01 and 0.05 and report the performance of the proposed method in Fig. 7. As shown in Fig. 7, for all three cases the algorithm renders the highest F_1 score when $\Delta\delta$ is close to 0.02.

Through the various experiments for parameters tuning, we also notice that although properly setting the parameters can achieve better performance, the performance of the proposed is generally not sensitive to the changes of the parameters within certain range.

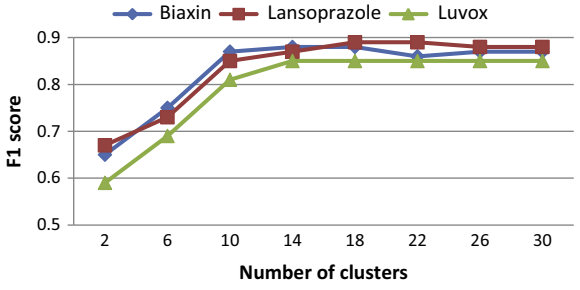

Fig. 6. ADR post identification accuracy using different number of clusters.

4.3. The performance evaluation

In this section, we compare the proposed approach with four benchmark methods (EAT, PNLH, ACTC and Laplacian SVM). The first three benchmark methods discussed in Section 2.2 are comparable with our approach in the following ways. First, they are independent from the classifier implemented. In addition, they employ the common idea of enlarging positive training examples during the learning process. In order to compare the performance of our approach with supervised classification methods, we utilize Laplacian SVM as the fourth benchmark method. Laplacian SVM builds upon the standard SVM framework. Researchers found that Laplacian SVM trained on labeled and unlabeled radiology reports significantly outperformed supervised SVMs [39].

Since the parameters involved in the four benchmark techniques greatly depend on the training corpus, and different settings of the parameter values may result in significant variation in performance. We first tune the parameters using the training set to capture the corpus-dependent aspect of each method. To ensure a fair comparison, we fine-tune the four benchmark methods according to their respective suggested guidelines [31,25,32,39]. In Table 3, we summarize the parameter values selected for the five techniques under examinations.

Due to the fact that the number of ADR posts is much less than those of the non-ADR posts found in the discussion threads, we examine the performance sensitivity of the different techniques to the size of the available positive examples. For all five methods, we conduct experiments using different proportion of the labeled ADR posts in the positive test data, ranging from 15% to 40%. The remaining test data in each corpus are used as unlabeled examples. For each of the three corpora, we generate 20 test sets via repeating the sampling process 20 times. The final performance measure is obtained by averaging the scores from all 20 test runs.

Tables 4–6 show the performance of the five methods with the percentage of the positive examples varies from 15% to 40%. Each column denotes the mean F_1 score on positive examples (ADR posts) of each method. We boldface the best classifier for each dataset. The last column denotes the improvement to the best performer in the benchmark methods. The paired-sample Wilcoxon signed-rank test was applied to assess the statistical significance with respect to the best benchmark method. For all three tables, we use the notation ** to indicate significance at $p < 0.01$; and *

Table 3
Summary of parameters tuning results.

Technique	Parameters index		Values
Our approach	T	The number of topics	35
	δ	The relevance threshold	0.55
	k	The number of clusters	18
	$\Delta\delta$	The increment of the relevance threshold	0.02
EAT	λ	The distance threshold for clustering	1.0
	τ	The minimum number of documents in an informative cluster	12
	ε	The parameter to facilitate the classifier construction process	10
PNLH	n	The number of top representative features	3700
	θ	The feature strength threshold	Self-tuning
	ϕ	The average positive referencing power for all positive cases	Self-tuning
ACTC	Maxlter	The number of iterations	60
	p	Percentage of unlabeled examples used in each iteration	0.5
Lapacian SVM	σ	Gaussian kernel parameter	Self-tuning
	p	Manifold estimation parameter	1
	γ_A, γ_I	The regularization parameters that control the smoothness of the separating hyperplane with respect to the ambient and intrinsic spaces	Self-tuning

to indicate significance at $p < 0.05$; unmarked results indicate no significant difference.

The p-values of the significance test results in Tables 4–6 show that our approach significantly outperform the benchmark techniques in most cases. In addition, we find that the performance difference is more significant when the size of the initial positive example is small. We believe the reason is that the proposed labeling heuristic is sufficiently smooth with respect to the intrinsic structure revealed by the labeled and unlabeled examples.

We highlight the strengths of the proposed approach as follows. First, we apply the LDA model to represent posts (documents) in the topic space instead of in the term space over the corpus. Such representation can effectively reduce the dimension of the texts while maintaining the discriminative power of the original representation. Second, the proposed labeling heuristic is probabilistic, allowing the incorporation of uncertainty into the label propagation process. Thus we can estimate how close an unlabeled post is to the labeled positive dataset according to the labeling confidence (probability). The labeling confidence measure is used to identify candidates for inclusion in the training set during the augmentation process. This is especially important when the size of the initially labeled positive example is small, as we do not know the distribution of the positive examples precisely in most problem situations [30]. As this is a common problem found when analyzing big data, the method can serve as an effective tool for social media analytics. Third, the proposed classifier construction process provides a mechanism that improves the separation of the positive and negative classes, thus prevents the labeling heuristic from suffering performance degradation caused by extracting too many noisy data.

4.4. Analysis of the detected ADRs

Another potential use of our approach is to gain better understanding of the common vocabularies found in consumer ADRs related messages in social media. We analyze the extracted topics from the posts filtered by our approach. Six topic examples, two for each drug, extracted from the Biaxin, Lansoprazole, and Luvox datasets are shown in Table 7, where each topic is related to a particular ADR.

The six topics shown on the top half of Table 7 were each represented by the top 10 topic words generated from the positive examples using the LDA model. On the bottom half of Table 7, we list the seven most common ADRs in adults taking the drug over the course of one year from the FDA online drug library.[2] As can be seen from the table, the words within each extracted topic are quite informative and coherent. For all three drugs, the top 10 topic words from each topic correspond well with the symptom in the documented ADRs. For example, the first topic of the Biaxin dataset is closely related to the adverse reaction "diarrhea", whereas the second topic is likely related to the side effect "nausea".

Comparing the extracted topics and the documented adverse reactions, we found the consumer health expressions are very diverse. These consumer health expressions have practical implications for the understanding of ADRs related self-disclosing health information; they include symptoms, mental status, behaviors, and help needed. The extracted topic words can not only improve the accuracy in identifying ADR posts on social media, but also boost effective communication, health information seeking, and ultimately informed decision making in post-marketing surveillance.

Table 4
The pairwise comparison of the F_1 scores (on ADR posts) with varying percentage of positive test examples for Biaxin data.

% Of positive examples	Benchmarks				Our approach	
	EAT (%)	PNLH (%)	ACTC (%)	Lapacian SVM (%)	Mean (%)	Improvement (%)
15	39.67	43.54	64.57	35.12	70.71	+6.14**
20	44.98	51.34	70.31	43.66	75.62	+5.31**
25	54.89	55.61	75.17	47.54	78.94	+3.77**
30	57.78	60.76	79.87	49.83	82.73	+2.86**
35	62.37	66.21	83.19	54.49	85.13	+1.94*
40	70.35	75.32	86.24	60.32	89.43	+3.19*

Table 5
The pairwise comparison of the F_1 scores (on ADR posts) with varying percentage of positive test examples for Lansoprazole data.

% Of positive examples	Benchmarks				Our approach	
	EAT (%)	PNLH (%)	ACTC (%)	Lapacian SVM (%)	Mean (%)	Improvement (%)
15	41.53	48.44	65.76	39.75	73.26	+7.50**
20	47.36	53.65	72.43	43.87	78.51	+6.08**
25	56.87	58.16	78.13	50.66	81.07	+2.94**
30	63.19	68.32	80.89	63.97	84.29	+3.40**
35	65.94	64.73	82.97	65.64	86.13	+3.16*
40	73.08	72.89	90.17	69.31	89.21	−0.96

Table 6
The pairwise comparison of the F_1 scores (on ADR posts) with varying percentage of positive test examples for Luvox data.

% Of positive examples	Benchmarks				Our Approach	
	EAT (%)	PNLH (%)	ACTC (%)	Lapacian SVM (%)	Mean (%)	Improvement (%)
15	40.24	43.63	69.33	37.39	71.37	+2.04**
20	45.29	52.77	73.67	43.55	77.96	+4.29**
25	57.31	58.89	77.51	54.84	80.38	+2.87**
30	62.07	67.24	80.29	57.93	82.51	+2.22**
35	68.52	72.53	82.97	61.53	83.98	+1.01*
40	74.29	79.59	85.56	63.54	85.43	−0.13

[2] http://www.accessdata.fda.gov/scripts/cder/drugsatfda.

Table 7
The comparison of extracted topics vs. documented adverse reactions.

Biaxin		Lansoprazole		Luvox	
Extracted topic examples					
Diarrhea	Nausea	Abdominal	Constipation	Headache	Asthenia
Diarrhoea	Sickness	Abdomen	Stool	Head	Fatigue
Dysentery	Stomach	Pain	Difficulty	Depression	Strength
Stool	Retch	Bellyache	Irregularity	Pain	Risk
Bowel	Upset	Stmachache	Difficult	Anxiety	Heart
Movement	Discomfort	Acute	Astriction	Help	Anxiety
Watery	Pain	Bloated	Intestinal	Suffer	Dizziness
Loose	Food	Swelling	Bowel	Suicide	Panic
Running	Vomit	Belly	Evacuation	Sick	Weak
Liquid	Queasiness	Question	Depression	Disorder	Somnolence
Documented ADRs					
Diarrhea, nausea, abnormal taste, dyspepsia, abdominal pain/ discomfort, headache		Abdominal pain, constipation, diarrhea, nausea, abdomen enlarged, asthenia		Headache, asthenia, nausea, chest pain, palpitation, diarrhea vasodilatation, hypertension	

5. Conclusions and future work

In this paper, we propose a novel framework for tackling the problem of filtering big data from social media in general, and the application to consumer ADR messages identification in specific. The framework contains three important components: the dimension reduction mechanism, the automatic augmentation of the training data, and the resulting classifier that can effectively extract consumer ADR related posts from health related social media.

We introduce the application of the LDA model as a dimension reduction technique that can be applied to various high-dimensional problems that social media analytics facing. We model the learning process as a partially supervised classification problem aided by a method for retrieving relevant unlabeled posts with the help of the LDA model. Through this method, we can automatically augment the training data (i.e., reliable positive and negative examples), thus build a more robust classifier for consumer ADR message identification in social media. An important contribution of this research is that the proposed approach can characterize the diversity of the consumer ADR posts in a low-dimensional feature space, in the meanwhile avoids performance degradation when augmenting the training data.

We validate the proposed method empirically using data collected from three Web forums in the Drugs section of MedHelp. First, we conduct experiments to determine the appropriate parameter settings. Through the parameter tuning process, we observe that the performance of the proposed method is generally not sensitive to the changes of the parameter values within a certain range. Second, we compare the ADR post filtering capability of the proposed method with those of four benchmark methods (EAT, PNLH, ACTC and Laplacian SVM). The empirical evaluations of the results from the three test beds (Biaxin, Lansoprazole, and Luvox) suggest that the proposed method generally outperforms the benchmark methods and exhibits more stable performance than its counterparts, especial when the available posts about certain ADRs are scarce. The outcome from this study could be used as input to an early warning system for detection of new ADRs.

For future research, we plan to extend our study in two major directions to address the current limitations. First, the current study only considers data collected from a particular social media platform, the Web forums. The performance of the proposed method using data collected from other social media platforms such as tweets or blogs has not been confirmed. Tweets and blogs are different in nature from forum posts, for example tweets are much shorter than forum messages. Hence, for our future research, we intend to explore the applicability of the proposed mechanism on other social media (Twitter, Facebook, etc.) and investigate on incorporating expert knowledge to guide the topic model learning process. Second, the topics of the consumer ADR data among different type of drugs can be quite dissimilar. Therefore, in order to maintain good classification performance, we will need to retrain the model by going through the whole model building process stating form crawling, parsing, and tagging the posts for each type of drug. However, such data-labeling process can be very time consuming and costly, let alone the time and effort needed to retrain the model. In order to reduce the effort needed for annotating consumer ADR messages for each different drug, we plan to extend the current method by incorporating the concept of transfer learning model [40], a data mining technique that allows for knowledge transfer from one learned classification domain to another domain. In other words, if successful, the knowledge we gained from learning the classification of consumer ADR messages of an antidepressant, e.g. Luvox, can be reused to train the classification model of consumer ADR messages of other type of drugs.

Acknowledgments

This work was partly supported by the Natural Science Foundation of China (Nos. 71301172, 71171186, 71301175, and 61272389) and Social Science Foundation of China (No. 13AXW010).

References

[1] Bruno HS, Bruce MP. Detection, verification, and quantification of adverse drug reactions. BMJ 2004;329:44–7.
[2] van Grootheest K, de Graaf L, de Jong-van den Berg LTW. Consumer adverse drug reaction reporting – a new step in pharmacovigilance? Drug Saf 2003;26:211–7.
[3] van Hunsel F, Talsma A, van Puijenbroek E, de Jong-van den Berg L, van Grootheest K. The proportion of patient reports of suspected ADRs to signal detection in the Netherlands: case-control study. Pharmacoepidemiol Drug Saf 2011;20:286–91.
[4] de Langen J, van Hunsel F, Passier A, de Jong-van den Berg L, van Grootheest K. Adverse drug reaction reporting by patients in the Netherlands – three years of experience. Drug Saf 2008;31:515–24.
[5] Fernandopulle RBM, Weerasuriya K. What can consumer adverse drug reaction reporting add to existing health professional-based systems? Focus on the developing world. Drug Saf 2003;26:219–25.
[6] Yang M, Li Y, Kiang M. Uncovering social media data public health surveillance. In: Proceedings of the 15th Pacific-Asia conference on information system, Brisbane; 2011. Paper 218.
[7] Fichman RG, Kohli R, Krishnan R. The role of information systems in healthcare: current research and future trends. Inform Syst Res 2011;22:419–28.
[8] Kane GC, Fichman RG, Gallaugher J, Glaser J. Community relations 2.0. Harvard Business Rev 2009;87:45–50.
[9] van Hunsel F, van der Welle C, Passier A, van Puijenbroek E, van Grootheest K. Motives for reporting adverse drug reactions by patient-reporters in the Netherlands. Eur J Clin Pharmacol 2010;66:1143–50.

[10] Yan P, Zeng D. Syndromic surveillance systems. Annu Rev Inform Sci Technol 2008;42:425–95.
[11] Abou Taam M, Rossard C, Cantaloube L, Bouscaren N, Roche G, Pochard L, et al. Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator®) withdrawal in France. J Clin Pharm Ther 2014;39:53–5.
[12] Leaman R, 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. In: Proceedings of the 2010 workshop on biomedical natural language processing. Association for Computational Linguistics, Uppsala, Sweden; 2010. p. 117–25.
[13] Yang CC, Yang HD, Jiang L, Zhang M. Social media mining for drug safety signal detection. In: Proceedings of the 2012 international workshop on smart health and wellbeing; 2012. p. 33–40.
[14] Lu HM, Zeng D, Trujillo L, Komatsu K, Chen H. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. J Biomed Inform 2008;41:340–56.
[15] Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. J Am Med Inform Assoc 2011;18:631–8.
[16] Zhang YL, Dang Y, Chen HC, Thurmond M, Larson C. Automatic online news monitoring and classification for syndromic surveillance. Decis Support Syst 2009;47:508–17.
[17] Torii M, Yin LL, Nguyen T, Mazumdar CT, Liu HF, Hartley DM, et al. An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. Int J Med Informatics 2011;80:56–66.
[18] Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Inter Res 2013;15:16.
[19] Denecke K, Nejdl W. How valuable is medical social media data? Content analysis of the medical web. Inform Sci 2009;179:1870–80.
[20] Huh J, Yetisgen-Yildiz M, Pratt W. Text classification for assisting moderators in online health communities. J Biomed Inform 2013;46:998–1005.
[21] O'Grady L, Wathen CN, Charnaw-Burger J, Betel L, Shachak A, Luke R, et al. The use of tags and tag clouds to discern credible content in online health message forums. Int J Med Informatics 2012;81:36–44.
[22] Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. In: Proceedings of AMIA symposium; 2011. p. 217–26.
[23] Lewis DD, Yang YM, Rose TG, Li F. RCV1: a new benchmark collection for text categorization research. J Mach Learn Res 2004;5:361–97.
[24] Chau M, Chen HC. Comparison of three vertical search spiders. Computer 2003;36:56–62.
[25] Fung GPC, Yu JX, Lu HJ, Yu PS. Text classification without negative examples revisit. IEEE Trans Knowl Data Eng 2006;18:6–20.
[26] Zhu X. Semi-supervised learning literature survey; 2005. <http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html>.
[27] Yu HJ, Han JW, Chang KCC. PEBL: web page classification without negative examples. IEEE Trans Knowl Data Eng 2004;16:70–81.
[28] Li X, Liu B. Learning to classify texts using positive and unlabeled data. In: Proceedings of the 18th international joint conference on artificial intelligence. Morgan Kaufmann Publishers Inc., Acapulco, Mexico; 2003. p. 587–92.
[29] Wei CP, Chen HC, Cheng TH. Effective spam filtering: a single-class learning and ensemble approach. Decis Support Syst 2008;45:491–503.
[30] Zhou K, Xue GR, Yang Q, Yu Y. Learning with positive and unlabeled examples using topic-sensitive PLSA. IEEE Trans Knowl Data Eng 2010;22:46–58.
[31] Ko HM, Lam W. A new approach for semi-supervised online news classification. In: Shimojo S, Ichii S, Ling TW, Song KH, editors. Web and communication technologies and internet –related social issues – Hsi 2005. Springer-Verlag Berlin Heidelberg; 2005. p. 238–47.
[32] Zhang X, Xiao W. Clustering based two-stage text classification requiring minimal training data. Comput Sci Inform Syst 2012;9:1627–43.
[33] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform 2012;45:885–92.
[34] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.
[35] Griffiths TL, Steyvers M. Finding scientific topics. In: Proceedings of the national academy of sciences of the United States of America; 2004. p. 5228–35.
[36] Cohen I, Cozman FG, Sebe N, Cirelo MC, Huang TS. Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. IEEE Trans Pattern Anal Mach Intell 2004;26:1553–67.
[37] Steyvers M, Griffiths T. Probabilistic topic models. Handbook Latent Semant Anal 2007;9:424–40.
[38] Minka T, Estimating a Dirichlet distribution, technical report, MIT; 2003.
[39] Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform 2013;46:869–75.
[40] Pan SJ, Yang QA. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345–59.

[2] http://www.accessdata.fda.gov/scripts/cder/drugsatfda.