Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2013 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

6-18-2013

# Identification of Consumer Adverse Drug Reaction Messages on Social Media

Ming Yang
*Central University of Finance and Economics*, yangming@cufe.edu.cn

Xiaodi Wang
*Central University of Finance and Economics*, wangxiaodi8@gmail.com

Melody Kiang
*California State University - Long Beach*, mkiang@csulb.edu

# IDENTIFICATION OF CONSUMER ADVERSE DRUG REACTION MESSAGES ON SOCIAL MEDIA

Ming Yang, School of Information, Central University of Finance and Economics, Beijing, China, yangming@cufe.edu.cn

Xiaodi Wang, School of Statistics, Central University of Finance and Economics, Beijing, China, wangxiaodi8@gmail.com

Melody Kiang, Department of Information Systems, College of Business Administration, California State University, Long Beach, United States, mkiang@csulb.edu

## Abstract

*The prevalence of social media has resulted in spikes of data on the Internet which can have potential use to assist in many aspects of human life. One prospective use of the data is in the development of an early warning system to monitor consumer Adverse Drug Reactions (ADRs). The direct reporting of ADRs by consumers is playing an increasingly important role in the world of pharmacovigilance. Social media provides patients a platform to exchange their experiences regarding the use of certain drugs. However, the messages posted on those social media networks contain both ADR related messages (positive examples) and non-ADR related messages (negative examples). In this paper, we integrate text mining and partially supervised learning methods to automatically extract and classify messages posted on social media networks into positive and negative examples. Our findings can provide managerial insights into how social media analytics can improve not only postmarketing surveillance, but also other problem domains where large quantity of user-generated content is available.*

*Keywords: Postmarketing surveillance, Consumer ADRs identification, Social media analytics, Text mining*

# 1 INTRODUCTION

With the advent of Web 2.0, the past few years have witnessed the rapid rise of social media, such as Web forums, blogs, micro-blogs, wikis and social network Web sites. The popularity of social media networks has resulted in an abundance of data on the Internet and awaiting for data analytics to discover their potential use. Public health officials have been looking closely at social media networks for possible identification of outbreaks of contagious diseases. Studies have shown that before the outbreak of 2009 H1N1 pandemic, there was a surge of activity on queries from Google users for keywords such as "muscle aches" and "thermometer". According to a 2011 study, tracking social media before the swine flu outbreak might have alerted the public health official weeks before it reached epidemic levels (Erin, 2012). In this study, we explore the possibility of using messages posted on social media networks for identification of potential adverse drug reactions.

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 (Bruno & Bruce, 2004). 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 (van Grootheest et al., 2003).

Typically, pharmacovigilance systems rely on the reporting by physicians and pharmacists, not directly from 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 to be 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 (van Grootheest et al., 2003). 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 (van Hunsel et al., 2011). Study has shown that consumer reporting of ADRs contributes significantly to a reliable pharmacovigilance ( Blenkinsopp, 2007; de Langen et al., 2008). However, not all countries accept consumer reports, especially for developing countries where around 80% of the global population lives (Fernandopulle & Weerasuriya, 2003). Also a considerable time lag still exists in recognition of serious ADRs using the consumer reporting. Hence, there is a need for a different approach to the existing pharmacovigilance.

The popularity of social media has resulted in a new way of communication among people that shares the same interests or experiences. Web forum is one of the most popular medium of providing fresh and fast information on the Web. Many posts shared on leading social network sites such as Facebook and Twitter were taken from Web forums. Social media provides patients a platform to exchange their experiences with drugs (Fox & Jones, 2009). Moreover, social media constitutes a significant part of the online search results for information about health and medical matters (Yang et al. 2011). Van Hunsel et al. (2010) 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 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 (Yan & Zeng, 2008). However, detecting "whispers of useful information in a howling hurricane of noise" is a huge challenge and better filters are needed to extract meaning from the "blizzard of buzz" (Woolridge, 2011).

In this paper, we explore the use of text and data mining techniques geared towards social media analytics to mine the alternative consumer ADR data sources: social media. 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 volumes of unlabeled content on social media. Our research endeavours 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 limited number of 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 two 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 reviews related works in social media analytics and their applications of text mining and data mining techniques. Section 3 summarizes current research gaps and the objectives of our study. In section 4, we present the methodology implemented that includes both text mining and data mining techniques and the data collection process. Section 5 evaluates the performance of the proposed system using datasets collected from two Web forums pertaining to consumer drug experiences. The paper is concluded with a summary of our contributions and suggestions for future research directions.

# 2 RELATED WORK

In this section, we review the literature on social media analytics and discuss the recent development and application of text mining technique to user-generated content on social media networks. A common challenge faced by social media analytics is the need for a computerized mechanism that can automatically filter and classify messages into positive and negative examples. We explore related works in social media analytics, text mining, and machine learning fields. We review the premises and limitations of prior work in approaching problems with majority of unlabeled data.

## 2.1 Social Media Analytics for Healthcare Management

Social media analytics deal with developing and evaluating informatics tools and frameworks to measure the activities within social media networks from around the Web (Yang et al., 2011). Data on conversations, engagement, sentiment, influence, and other specific attributes can then be collected, monitored, analyzed, summarized, and visualized. A growing body of research on social media analytics has found applications in healthcare related tasks. For instance, Denecke & Nejdl (2009) 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 various health related social media. Boulos et al. (2010) evaluated a number of emerging technologies and tools that can be used to exploit these inherent social Web features in real or near-real time and harness them for public health, environmental and national security surveillance purposes. Based on text mining and social network analysis, Corley et al. (2010) proposed an approach that can identify online communities for targeted public health communications to assure wide dissemination of pertinent information.

Consumer posting of suspicious ADRs in social media has the potential to increase the understanding and provide early warning of possible harms of a medicine. However, due to the lack of an automatic mechanism to filter and classify ADR related posts from non-ADR ones, it has substantially hampered the potential advancement in the development of an early warning system for ADRs based on user-generated content on social media networks. Moreover, although some countries allow for consumer direct reporting of ADRs, people are more willing to share their experiences by posting on forums than writing up a formal report for submission to an official reporting system. Often times the

reactions experienced by the consumers may not be serious enough to constitute a report right away. However, when a good number of consumers post about the same reaction of a certain drug, it may suggest the need for further investigation. The timely identification of potential ADRs can mean potential life and cost savings and we believe this can be achieved through the analysis of data from social media networks.

## 2.2 Text Mining in Healthcare

Text mining techniques to user-generated content in social media has recently gained considerable attention. Text mining has found applications in a wide spectrum of problem domains, including information extraction, topic tracking, summarization, categorization, concept linkage, clustering, information visualization, content filtering, and prediction (Gopal et al., 2011).

Extant studies have applied text mining to various aspects of public healthcare problems. For example, Lu et al. (2008) proposed a new ontology-enhanced approach for classifying free-text chief complaints (CCs) from the emergency department. In order to provide adequate support for processing CCs recorded in non-English languages, Lu et al. (2009) developed a multilingual CC classification system leveraging a Chinese-English translation module and an existing English CC classification approach. Zhang et al. (2009) developed a text mining framework for automatic online news monitoring and classification to track the emergence of health epidemics.

However, the content of medical social media data can be diverse due to the background of the author, and the source or the topic. Diversity in the context presents opportunities and challenges for syndromic surveillance due to the requirement on fast analysis of relevant Web data. With the attempt to create a "gold standard" data set to test how accurately disease outbreak information extraction systems can identify the semantics of disease outbreak events, Conway et al. (2010) developed an annotation scheme for identifying infectious disease outbreak events in news texts. In order to create a technique for precise and task-driven Web data collections which are rich in content that meets specific requirements, Fu et al. (2012) proposed a novel focused crawler that incorporates topic and sentiment information as well as graph-based tunnelling mechanism for enhanced collection of opinion-rich web content regarding post-marketing drug surveillance.

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 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 (Lewis et al., 2004).

## 2.3 Learning with Positive and Unlabeled Data

Supervised learning algorithms require high-quality labeled training data in order to construct an accurate classifier. However, messages related to consumer ADRs are usually scarce and highly distributed in social media networks. 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 is of varying quality, and difficult to locate on the Web (Chau and Chen, 2003). Finally, due to the dynamically changing environment of the social media networks, the labeled training data may become outdated quickly.

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 (Zhu, 2005). Fung et al. (2006) 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. In many information retrieval applications, positive

examples refer to the data points that are of interest to the researchers in a binary classification problem. In this research, we assign the messages containing consumer ADRs as positive examples and non-consumer ADR messages as negative examples to train the classifiers.

The positive class is usually more specific than the negative class (Zhou et al., 2010). Hence, it is possible to identify more potentially positive examples from the unlabeled data through exploiting the inherent structures in the set of positive examples. Ko and Lam (2005) 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 for 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 (Fung et al., 2005). In order to solve this problem, Fung et al. (2006) 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 this assumption fails, that is, when the available positive features are not representative of the true positive class, the performance of this method may deteriorate.

# 3 RESEACH GAPS AND RESEARCH QUESTIONS

Based on our review of the related works, we have identified several important research gaps. Due to the fact that customer ADR posts on social media are highly distributed and non-ADR content is very diverse in topics, it is costly and time consuming to label a large number of training data for a supervised learning algorithm to identify consumer ADR posts from unlabeled dataset. One way to overcome such difficulties is to dynamically augment the training data through a partially supervised learning (PSL) process.

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 the significance of each others' core vocabulary (Fung et al., 2006). Eventually, it is difficult to extract reliable positive and negative examples based on the core vocabulary. Thus existing PSL techniques may have a limited ability to capture the diversity of the positive class in this situation. The deficiency of existing techniques for identifying consumer ADRs in Web forum, coupled with the challenges associated with building more effective ones warrants the use of new guidelines to help inform future system development. In particular, this study seeks to answer the following research questions:

1) How can we use text features to characterize the diversity of the consumer ADR posts?

2) Given the special characteristics of consumer ADR posts on social media networks, can our approach provide further identification power?

# 4 METHODOLOGY

Since no negative example is given explicitly in PSL, 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 examples (Zhu, 2005). A labeling heuristic must be designed which is sufficiently smooth with respect to the intrinsic structure revealed by labeled and unlabeled points. 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 PSL approach to monitor content containing negative sentiments regarding various drugs and medicines at targeted Web forums. Such sentiments could be important indicators of potential adverse drug reactions (Chee et al. 2011). The proposed approach has four components – data collection, feature acquisition, consensus detection, and classifier construction (as shown in Figure 1). These components are described in the following sub-sections.

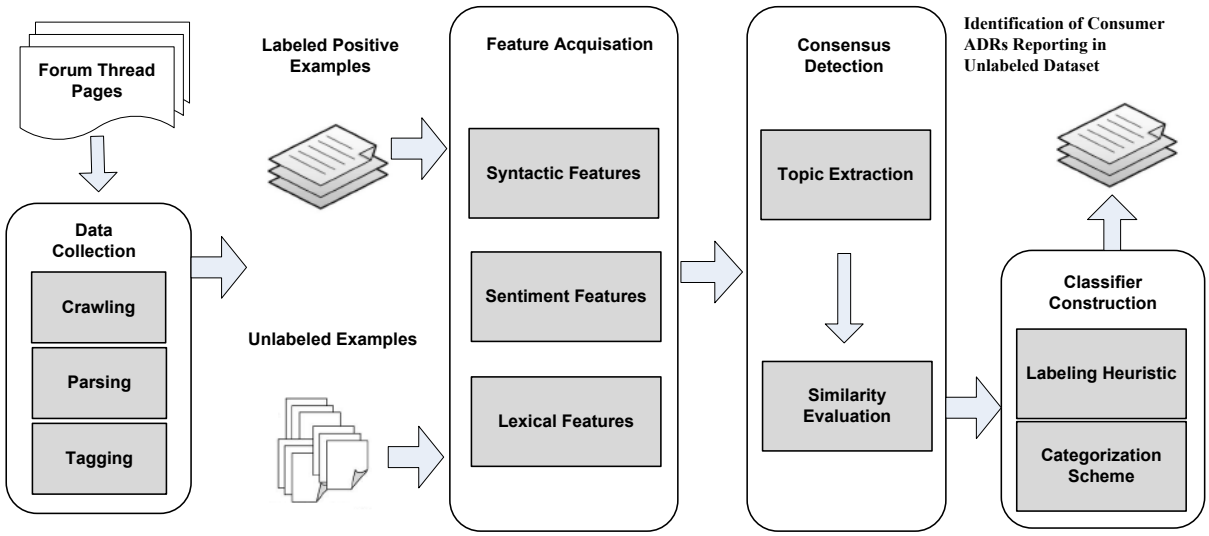


*Figure 1. A methodology for identification of consumer ADR messages on Web forums*

## 4.1 Data Collection

We begin with the crawling of Web forums to gather patient posts from social media networks. Parsing programs are then developed to convert the posts from the raw HTML pages and store it in a relational database. We extract the username, date, time of post, and the discussion text. Finally, domain experts are recruited to tag the collected messages independently. They must assess whether an ADR is being discussed, how critical the ADR is, and what drug the ADR relates to. Further annotations can also be added (e.g., what the implications of the ADR are; why a syndrome should not be regarded as an ADR). After the data collection stage, we construct the labeled positive example set (a small number of labeled consumer ADR messages) and the unlabeled example set (a mixture of consumer ADR messages and non-ADR messages).

## 4.2 Feature Acquisition

In order to reduce the complexity of text documents (collected by the crawling programs) and make them easier to handle, full text documents have to be transformed to document vectors which describe the contents. The discourse in Web forums can be manifested in the form of various information types, including topics, events, opinions, emotions, styles, and interactions. Converting the information types into a feature vector that makes its most salient features available is an important part of text mining (Abbasi and Chen, 2008). In order to obtain such an enormous mass of information, we incorporated the three feature categories based on a review of previous studies, i.e., syntactic, semantic, and sentiment features. The complete list of the features used in our study is presented in Table 1.

The syntactic features mainly contain punctuation marks, function words, and part-of-speech (POS)

tags n-grams. To form the syntactic feature set, we adopt the set of 149 function words suggested by (Zheng et al., 2006) and the 8 punctuation words suggested by Baayen et al. (1996). Moreover, we employ POS tags uni-, bi-, and tri-grams to capture the writing style at the sentence level. Semantic features encompassed the use of bag-of-words, noun phrases, and named entities. In this study, we adopt all the three types of semantic features. Zhang et al. (2009) showed that the combined three semantic feature groups improved the performance of online news classification. Sentiment features refer to words and/or phrases used to determine the overall sentiment of a post. Abbasi et al. (2008) had suggested the use of affect lexicons as sentiment features to capture large volume of terms for expressing emotions. The affect term list shown in Table 1 is manually created by the domain experts following the process suggested by (Subasic and Huettner, 2001).

| Features | Group | Quantity | Description / examples |
|---|---|---|---|
| Syntactic | Function words | 149 | Frequency of function words (e.g., of, for, to) |
| | Punctuation | 8 | Occurrence of punctuation marks (e.g., !, : ; ?) |
| | POS tag n-grams | Varies | Part-of-speech tag n-grams (e.g., NNP, NNP JJ) |
| Semantic | Bag-of-words | Varies | All words except function words |
| | Noun phrases | 200 | Syndrome, side effect lexicons |
| | Named entity | 37 | Medical terminology, drug names, diseases names |
| Sentiment | Affect lexicons | 157 | Worry, fear, anger, etc. |

*Table 1. The text features adopted in this research*

## 4.3 Consensus Detection

We define consensus as broad declarative statements reflecting a majority of certain group of consumers' opinions. Consensus in the communication can provide great impact on individual's causal judgment when the communication involves negative information (Conway et al., 1990). However, in Web forums the consumer ADR posts which are of interest to the concern of FDA and pharmaceutical companies usually consist of diverse syndrome topics and are highly distributed. Hence, it is challenging to detect the community's consensus for further analysis. In the proposed method, we detect the consensus of the consumer via two steps, i.e., topic extraction and similarity evaluation.

To extract topics from the labeled positive training set S, we first ask domain experts to identify representative messages as seeds from set S. Then, using the seeds, we perform text clustering on the entire set S. For each element $s_i$ in the set S, we calculate its similarity with the centroid of each cluster. If the similarity is closer than a threshold $r$, $s_i$ will be assigned to that cluster. After the clustering process, the clusters contain more than $\tau$ documents are considered informative clusters. For each identified informative cluster, we apply a Mutual Information (MI)-based noun phrase extractor, Arizona Noun Phraser (Tolle and Chen, 2000), to extract major terms that can represent the topic.

Since the number of labeled positive examples is much smaller than that of the unlabeled examples, in order to detect the consensus of the topics extracted from the informative clusters, we need to evaluate the similarity of all messages in the unlabeled set P for each topic extracted from the informative clusters. If the similarity of a message in P with at least one topic pattern of the informative cluster is greater than a given threshold $\varepsilon$, the message is selected and introduced into the available repertoire A. At the end of the process, the available repertoire A is merged with the original positive training set S, to form the final positive training set that contains consumer ADR posts with high level of consensus; while the rejected messages form the negative training set. Through this method, consumer ADRs with high level of consensus are identified; in the meantime more positive examples and negative examples for training can be obtained.

The similarity evaluation algorithm aims at selecting messages from unlabeled set (P) into either

available repertoire or negative examples set (N). It takes two inputs, unlabeled set (P) and informative clusters ( $S^M$ ). For each $S_j^M \in S^M$ , we let all messages in P be negative examples and those in $S_j^M$ be positive examples. Then two prototype vectors *x* and *y*, corresponding to positive and negative prototype respectively, are learned by the Rocchio algorithm (Li and Liu, 2003). For each $p_i \in P$ , we calculate its similarity with $S_j^M$ as follows:

$$\varphi(p_i, S_j^M) = \frac{p_i \cdot x}{\|p_i\| \cdot \|x\|} - \frac{p_i \cdot y}{\|p_i\| \cdot \|y\|} \quad (1)$$

The Rocchio algorithm is utilized to build a classifier using both the informative cluster and the unlabeled set. The classier is then applied to select more positive examples into available repertoire and delete them from unlabeled set. The idea is to identify more positive examples in unlabeled dataset in a localized manner. In terms of the similarity measure, $p_i$ maybe similar to both $S^M$ and $N_i$. We only extract $p_i$ which is significantly similar to either $S^M$ or $N_i$. The output of the similarity evaluation process is the available repertoire and the negative examples set.

## 4.4 Classifier Construction

Previous studies have found that the positive examples extracted through comparing the differences of the feature distributions between the positive class and the unlabeled dataset does not always guarantee all the extracted positive examples are reliable (Fung et al., 2006). Moreover, when too many labeled examples were extracted, it may cause over-fitting problem and may lead to performance degradation (Cohen et al., 2004). In order to solve these problems, we select reliable positive and negative examples to fit the distribution of the positive and negative classes, respectively. The process is controlled by iteratively running a classification scheme $C_m$. We choose Support Vector Machine (SVM) as the text classifier due to its popularity and superb performance in text classification (Fung et al., 2006). To guarantee the quality of positive examples in the available repertoire is in a stable state, we keep the change of the threshold $\Delta\varepsilon$ during *m*th iteration within a limited range. We use the following updating rule: $\Delta\varepsilon = e^{-m/\sigma}$, where *m* is the number of iteration, $\sigma$ is the parameter that controls its decay.

Figure 2 shows the steps of the classifier construction algorithm. To build classifier $C_m$, we let $A_m$ be the positive training set that consists of consumer ADR posts with high level of consensus, and let $N_m$ be the negative training set that consists of messages that could not match any topic extracted from the informative clusters. The classifier $C_m$ is selected based on its local maximum $F_1$ score (for positive class) on the dynamically updated training data (in Lines 8-13).

**Input:** $A_0$ (available repertoire) and $N_0$ (rejected messages set), *P* (unlabeled dataset), $C_0$ (a classifier built by $A_0$ as the positive set and $N_0$ as the negative set), and $F_0$ (the F-measure of $C_0$ for the positive class);

**Output:** $A_m$ (updated available repertoire) and $N_m$ (updated rejected messages set);

1. $A_m \leftarrow A_0; N_m \leftarrow N_0;$
2. Initialize $F_{\max} = F_0; m = 1;$
3. **repeat**
4. $\varepsilon = \varepsilon_0 * (1 + \Delta\varepsilon_m)$
5. Obtain *A* and *N* through the similarity evaluation;
6. $A_m \leftarrow A; N_m \leftarrow N;$
7. Construct classifier $C_m$ using $A_m$ as the positive training set and $N_m$ as the negative training set;
8. Let $F_m$ be the F-measure of $C_m$ for the positive class;

9. **if** $F_m \geq F_{\max}$ **then**
10. $F_{\max} = F_m$;
11. **end if**
12. $m$++;
13. **until** $F_{m-1} < F_{\max}$
14. **return** $A_m$ and $N_m$;

*Figure 2 The Classifier Construction Algorithm*

# 5 EVALUATION AND ANALYSIS OF RESULTS

In this section, we report our empirical evaluation of the integrated text mining and partially supervised learning approach for consumer ADR post identification. We first describe our experiment process, which includes the datasets, the experiment design, and the evaluation criteria. This is followed by an analysis of the results.

## 5.1 The Datasets

In this research, we collect online messages from two health related Web forums. Prozac is a drug used to treat major depressive disorder, bulimia nervosa (an eating disorder) obsessive-compulsive disorder, panic disorder, and premenstrual dysphoric disorder, and ProzacAwareness is a forum in Yahoo! Groups about Health & Wellness. SSRIsex is another forum in Yahoo! Groups for people with sexual side effects that began during Selective Serotonin Reuptake Inhibitors (SSRIs) antidepressants use, but have continued after quitting them. The online forums we use consist of 28,340 public health & Wellness Yahoo! Groups. Within these groups there are a total of 13,647,986 messages. These groups range from illness based support groups focusing on drugs and medications to health care related groups. The messages within these groups span 9 years and consist of hundreds of thousands of unique email addresses that we use as proxy for the user.

We first download the 2011 snapshot of safety alerts for human medical products from U.S. FDA website. To keep the study at a manageable size, we extract the top 200 most frequent ADR description keywords, which represent the most common ADRs. We then extract the top 900 threads from each forum, which contain the most mentions of these ADRs. Three independent medical domain experts were employed to conduct the thread tagging. The experts were required to read each thread assigned to them and decide whether the threads discuss ADRs. From the tagged threads, we further extracted ADR posts and non-ADR posts. We limited our test bed to the two particular groups, ProzacAwareness and SSRIsex. Table 2 shows summary statistics for the two test beds.

| Groups | ADR posts | Non-ADR posts | Authors | Average Lengths (Char.) |
|---|---|---|---|---|
| Prozac Awareness | 800 | 2,400 | 129 | 1,043 |
| SSRIsex | 800 | 2,400 | 136 | 1,104 |

*Table 2. The Data sources*

## 5.2 Experiment Setup

In order to assess the consumer ADR posts identification power of the proposed method, we compare its performance through two experiments. For each experiment, we randomly selected $x$ percent of

consumer ADR posts in the corresponding test bed and use these documents to form the positive examples. Here, *x* ranges from 20 to 40 in increments of 5. The remaining training documents in each test bed were regarded as unlabeled examples. We performed the experiment 30 times and report the average of the results. We implement the $F_1$ measure to evaluate the reliability of the extracted positive examples, and the $F_1$ measure is calculated as follows:

$$F_1(i) = \frac{2 \times \text{precision}(i) \times \text{recall}(i)}{\text{precision}(i) + \text{recall}(i)} \tag{2}$$

where

$$\text{precison}(i) = \frac{\text{number of correctly classified cases for class } i}{\text{total number of cases classified as class } i}$$

$$\text{recall}(i) = \frac{\text{number of correctly classified cases for class } i}{\text{total number of cases in class } i}$$

## 5.3 PSL vs. without PSL

In this section, we evaluate the effectiveness of our approach in classifying the unlabeled training examples using two popular classifiers: SVM and Naïve Bayes (NB). Given a set of positive examples and a set of unlabeled posts, for each classifier, we compare two cases: with PSL versus without PSL. For the former, the two classifiers employ labeled positive examples (*x%* of consumer ADR posts) and all the non-ADR posts in the corresponding test beds as training data. For the latter, the two classifiers utilize labeled positive examples (*x%* of consumer ADR posts) and part of the unlabeled examples (*1-x%* of consumer ADR posts and all the remaining non-ADR posts in the corresponding test bed) as training data.

| % | Without PSL | | With PSL | |
|---|---|---|---|---|
| | SVM | NB | SVM | NB |
| 20 | 24.69% | 27.57% | 64.74% | 61.49% |
| 25 | 29.03% | 32.17% | 69.77% | 66.81% |
| 30 | 39.76% | 39.43% | 73.52% | 70.36% |
| 35 | 53.48% | 50.64% | 80.94% | 78.84% |
| 40 | 64.87% | 61.59% | 88.94% | 85.63% |

*Table 3 Results from the ProzacAwareness group*

| **%** | Without PSL | | With PSL | |
|---|---|---|---|---|
| | SVM | NB | SVM | NB |
| 20 | 23.59% | 27.89% | 65.31% | 60.98% |
| 25 | 30.67% | 34.53% | 71.04% | 68.59% |
| 30 | 45.72% | 42.88% | 76.63% | 72.52% |
| 35 | 54.66% | 52.31% | 83.26% | 80.14% |
| 40 | 67.74% | 64.59% | 89.74% | 86.32% |

*Table 4 Results from the SSRIsex group*

Tables 3 and 4 show the experiment results of ProzacAwareness group and SSRIsex group, respectively. In each table, the first column gives percentage of the total number of positive examples utilized. The next two columns present the performance ($F_1$ score) of the two classifiers for consumer

ADR posts identification without PSL. And the last two columns show the performance of the two classifiers for consumer ADR posts identification using the proposed PSL approach as the labeling heuristic. As shown in the two tables, the classifiers that employ PSL for labeling the unlabeled training examples were more effective, especially when the size of positive training examples is small.

For each classifier, we compare its performance between with PSL vs. without PSL on the $F_1$ socre for identifying consumer ADR posts. We conduct 10 individual paired-sample Wilcoxon signed-rank test for the performance comparison on each test bed (*p*-values significant at alpha=0.01). We found that for each individual paired-sample Wilcoxon signed-rank test, the classifier using PSL as the labeling heuristic achieved significantly higher $F_1$ score than the same classifier without PSL across all sample sizes. Hence, the proposed PSL approach can boost the effectiveness of the classifiers in identifying consumer ADR posts, in particular when the proportion of labeled positive example is small.

## 5.4 Performance Comparison of Different Labeling Heuristics

In this section, we compare the proposed approach with two benchmark labeling heuristics (EAT and PNLH) as discussed in section 2.2. The three labeling heuristics are comparable in the following ways. First, they are independent from the classifier implemented. In addition, they employ the common idea of enlarging the positive training examples during the learning process. In this experiment, we chose SVM as the text classifier, since SVM outperformed NB in our previous experiments. As the implementations of EAT and PNLH are not publicly available, we implemented the two algorithms based on the descriptions in (Ko and Lam, 2005) and (Fung et al., 2006), respectively.

Another objective of this experiment is to compare the sensitivity of the three techniques to the size of the initial positive examples. Tables 5 and 6 show the performance of the three labeling heuristics with the size of the positive examples varies from 20% to 40% of consumer ADR posts in the corresponding corpus. We also calculate the *p*-values of the paired-sample Wilcoxon signed-rank test between each of the two benchmark labeling heuristics and our approach. The corresponding *p*-values are also reported in Tables 5 and 6.

| **Size of Initial Positive Examples** | **20%** | **25%** | **30%** | **35%** | **40%** |
|---|---|---|---|---|---|
| Our Approach | 64.74% | 69.77% | 73.52% | 80.94% | 88.94% |
| EAT | 43.79%** | 50.44%** | 53.17%** | 65.31%** | 81.79%* |
| PNLH | 52.19%** | 56.77%** | 66.87%** | 75.37%** | 81.08%** |

**p-values significant at alpha<0.05*

***p-values significant at alpha<0.01*

*Table 5 Results from the ProzacAwareness group*

| **Size of Initial Positive Examples** | **20%** | **25%** | **30%** | **35%** | **40%** |
|---|---|---|---|---|---|
| Our Approach | 65.31% | 71.04% | 76.63% | 83.26% | 89.74% |
| EAT | 43.19%** | 49.64%** | 56.14%** | 65.96%** | 81.77%* |
| PNLH | 52.24%** | 56.03%** | 64.28%** | 71.94%** | 79.31%** |

**p-values significant at alpha<0.05*

***p-values significant at alpha<0.01*

*Table 6 Results from the SSRIsex group*

The *p*-values of the significance test results in Tables 5 and 6 show that our approach significantly outperform the benchmark techniques in all cases. In addition, we find that the performance difference is more significant when the size of the initial positive examples is small. We believe the reason is that the proposed labeling heuristic is sufficiently smooth with respect to the intrinsic structure revealed by labeled and unlabeled examples.

We highlight the strengths of the proposed approach as follows. First, the combination of syntactic, semantic, and sentiment feature set is capable of characterizing the discourse structure of a message. Second, consensus detection is able to capture the diversity of the positive class so as to extract reliable positive and negative examples. This is especially important when the size of the initially labeled positive examples is small, as we do not know the distribution of the positive examples precisely in most problem situations (Zhou et al., 2010). Third, the classifier construction process provides a mechanism that improves the separation of the positive and negative classes, which prevents the labeling heuristic from suffering performance degradation caused by extracting too many labeled examples.

# 6 CONCLUSION AND FUTURE DIRECTIONS

In this research, we propose a social media monitoring system that combines text mining and a partially supervised learning algorithm to identify the consumer ADR messages in online health forums. 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 ADRs identification on social media. An important contribution of this research is that the proposed approach can dynamically capture the characteristics of the positive class with diverse topics, in the meanwhile avoids performance degradation while augmenting the training data. We validate the proposed method empirically using data collected from two Web forums in Yahoo! Groups related to public health & wellness. We find that the proposed approach generally outperforms the benchmark techniques and exhibits more stable performance, in particular, when existing knowledge about certain ADRs are scarce.

To the best of our knowledge, the experiments conducted in this study are the first to apply partially supervised learning algorithm to the domain of consumer ADRs identification in social media networks. The outcome from this study could be used as input to existing or new warning systems for early detection of new drug ADRs (Bate et al., 1998; Van Puijenbroek et al., 2003).

Although the preliminary findings of our study are encouraging, more research is needed to further validate the applicability of our method with drugs other than antidepressants. For future research, we will extend our study in two major directions to address the current limitations. First, the 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 alternative social media networks (Twitter, Facebook, etc.) and investigate on utilizing additional linguistic features of the text to improve the overall performance of the model. Second, the unique features 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 re-train the model by collecting a large number of labeled data 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 re-train 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 [Pan & Yang 2010], 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 antidepressants can be reused to train the classification model of consumer ADR messages of other type of drugs.

## Acknowledgement

This work was supported in part by the grants from the National Natural Science Foundation of China (No. 61272389) and Natural Science Foundation of Beijing (No. 4112053). The authors also acknowledge support from the planning project of philosophy and social science of Beijing (No. 12JGA014).

## References

Abbasi, A., Chen, H.C., Thoms, S., Fu, T.J. (2008). Affect analysis of web forums and blogs using correlation ensembles. IEEE Transactions on Knowledge and Data Engineering, 20(9), 1168-1180.

Abbasi, A., Chen, H.C. (2008). Cybergate: A design framwork and system for text analysis of computer-mediated communication. MIS Quarterly, 32(4), 811-837.

Baayen, H., van Halteren, H., Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3), 121–132.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. European Journal of Clinical Pharmacology, 54(4), 315 - 321.

Blenkinsopp, A., Wilkie, P., Wang, M., Routledge, P. A. (2007). Patient reporting of suspected adverse drug reactions: a review of published literature and international experience. British Journal of Clinical Pharmacology, 63(2), 148–156.

Erin, A. (2012). Google, Twitter being used to track flu outbreaks. Scripps Howard News Service. Available at: http://www.standard.net/stories/2012/11/15/google-twitter-being-used-track-flu-outbreaks

Boulos, M.N.K., Sanfilippo, A.P., Corley, C.D., Wheeler, S. (2010). Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. Computer Methods and Programs in Biomedicine, 100(1), 16-23.

Bruno, H.S., Bruce, M.P. (2004). Detection, verification, and quantification of adverse drug reactions. BMJ, 329(7456), 44-47.

Chau, M., Chen, H.C. (2003). Comparison of three vertical search spiders. IEEE Computer, 36(5), 56-62.

Chee, B. W., Berlin R., Schatz, B. (2011). Predicting adverse drug events from personal health messages. In the Proceedings of AMIA Symposium, p. 217-226.

Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S. (2004). Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(12), 1553-1567.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dinh, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K. (2008). BioCaster: Detecting public health rumors with a Web-based text mining system. Bioinformatics, 24(24), 2940-2941.

Conway, M., Kawazoe, A., Chanlekha, H., Collier, N. (2010). Developing a disease outbreak event corpus.Journal of Medical Internet Research, 12(3): e43.

Conway, M., Difazio, R., Bonneville, F. (1990). Consensus and causal attributions for negative affect. Journal of Social Psychology, 130(3), 375-384.

Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P. (2010). Text and Structural Data Mining of Influenza Mentions in Web and Social Media. International Journal of Environmental Research and Public Health, 7(2), 596-615.

De Langen, J., van Hunsel, F., Passier, A., de Jong-van den Berg, L., van Grootheest, K. (2008). Adverse drug reaction reporting by patients in the Netherlands - Three years of experience. Drug Safety, 31(6), 515-524.

Denecke, K., Nejdl, W. (2009). How valuable is medical social media data? Content analysis of the medical web. Information Sciences, 179(12), 1870-1880.

Fernandopulle R.B.M., Weerasuriya K. (2003). What can consumer adverse drug reaction reporting add to existing health professional-based systems? Focus on the developing world. Drug Safety, 26(4), 219-225.

Fu, TJ., Abbasi, A., Zeng, D., Chen, HC. (2012). Sentimental spidering: Leveraging opinion information in focused crawlers. ACM Transactions on Information Systems, 30(4), article 24.

Fung, G.P.C., Yu, J.X., Hongjun, L., Yu, P.S. (2005). Text classification without labeled negative documents. In Proceedings of the 21st International Conference on Data Engineering (IEEE Computer Society), p. 594-605, Japan, Tokyo.

Fung, G.P.C., Yu, J.X., Lu, H.J., Yu, P.S. (2006). Text classification without negative examples revisit. IEEE Transactions on Knowledge and Data Engineering, 18(1), 6-20.

Fox S, Jones S. (2009). The social life of health information. Washington DC: Pew Internet & American life project. Available at: http://www.pewinternet.org/ ~/media//Files/Reports/2009/ PIP_Health_2009.pdf

Gopal, R., Marsden, J.R., Vanthienen, J. (2011). Information mining - Reflections on recent advancements and the road ahead in data, text, and media mining. Decision Support Systems, 51(4), 727-731.

Ko, H.M., Lam, W. (2005). A new approach for semi-supervised online news classification, in: Shimojo, S., Ichii, S., Ling, T.W., Song, K.H. (Eds.), Web and Communication Technologies and Internet -Related Social Issues - Hsi 2005, p. 238-247, Springer-Verlag Berlin, Berlin.

Lewis, D.D., Yang, Y.M., Rose, T.G., Li, F. (2004). RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 5, 361-397.

Li, X., Liu, B. (2003). Learning to classify texts using positive and unlabeled data, In Proceedings of the 18th International Joint Conference on Artificial Intelligence, p. 587 – 592, Morgan Kaufmann, San Francisco.

Lu, H.M., Chen, H., Zeng, D., King, C.C., Shih, F.Y., Wu, T.S., Hsiao, J.Y. (2009). Multilingual chief complaint classification for syndromic surveillance: An experiment with Chinese chief complaints. International Journal of Medical Informatics, 78(5), 308-320.

Lu, H.M., Zeng, D., Trujillo, L., Komatsu, K., Chen, H. (2008). Ontology-enhanced automatic chief complaint classification for syndromic surveillance. Journal of Biomedical Informatics, 41(2), 340-356.

Pan, S., J., Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. IEEE Transactions on Fuzzy Systems, 9(4), 483 – 496.

Tolle, K.M., Chen, H.C. (2000). Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science, 51(4), 352-370.

Van Grootheest, K., de Graaf, L., de Jong-van den Berg, L.T.W. (2003). Consumer adverse drug reaction reporting - A new step in pharmacovigilance. Drug Safety, 26(4), 211-217.

Van Hunsel, F., Talsma, A., van Puijenbroek, E., de Jong-van den Berg, L., van Grootheest, K. (2011). The proportion of patient reports of suspected ADRs to signal detection in the Netherlands: case-control study. Pharmacoepidemiology and Drug Safety, 20(3), 286-291.

Van Hunsel, F., van der Welle, C., Passier, A., van Puijenbroek, E., van Grootheest, K. (2010). Motives for reporting adverse drug reactions by patient-reporters in the Netherlands. European Journal of Clinical Pharmacology, 66(11), 1143-1150.

Van Puijenbroek, Eugène P., Diemont W. L, van Grootheest K. (2003). Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions. Drug Safety, 26(5), 293-301.

Woolridge A. (2011). Social media provides huge opportunities, but will bring huge problems. Economist, 50.

Yan, P., Zeng, D. (2008). Syndromic surveillance systems. Annual Review of Information Science and Technology, 42, 425-495.

Yang, M., Kiang, M., Ku, Y., Chiu, C., Li, Y. (2011). Social Media Analytics for Radical Opinion Mining in Hate Group Web Forums. Journal of Homeland Security and Emergency Management, 8(1), article 38.

Yang, M, Li, Y., Kiang, M. (2011). Uncovering social media data public health surveillance. In the Proceedings of the 15th Pacific-Asia Conference on Information System, paper 218, Australia, Brisbane.

Zhang, Y.L., Dang, Y., Chen, H.C., Thurmond, M., Larson, C. (2009). Automatic online news monitoring and classification for syndromic surveillance. Decision Support Systems, 47(4), 508-517.

Zheng, R., Li, J.X., Chen, H.C., Huang, Z. (2006). A framework for authorship identification of Online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology, 57(3), 378-393.

Zhou, K., Xue, G.R., Yang, Q., Yu, Y. (2010). Learning with Positive and Unlabeled Examples Using Topic-Sensitive PLSA. IEEE Transactions on Knowledge and Data Engineering, 22(1), 46-58.

Zhu, X. (2005). Semi-Supervised Learning Literature Survey. Available at: http://pages.cs.wisc.edu/ ~ jerryzhu/research/ssl/semireview.html