Journal of Biomedical Informatics xxx (2015) xxx–xxx

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 xxxx

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 biosurveillance 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

Please cite this article in press as: Yang M et al. Filtering big data from social media – Building an early warning system for adverse drug reactions. J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.01.011