Journal of Biomedical Informatics 58 (2015) 268–279

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports

Xiao Liu [a,*], Hsinchun Chen [a,b]

[a] Department of Management Information Systems, The University of Arizona, Tucson, AZ, United States
[b] Tsinghua National Laboratory for Info. Science and Technology, Tsinghua University, Beijing, China

ARTICLE INFO

Article history:
Received 6 January 2015
Revised 20 October 2015
Accepted 21 October 2015
Available online 27 October 2015

Keywords:
Knowledge acquisition
Information search and retrieval
Health social media analytics
Adverse drug event extraction
Text mining
Pharmacovigilance

ABSTRACT

Social media offer insights of patients' medical problems such as drug side effects and treatment failures. Patient reports of adverse drug events from social media have great potential to improve current practice of pharmacovigilance. However, extracting patient adverse drug event reports from social media continues to be an important challenge for health informatics research. In this study, we develop a research framework with advanced natural language processing techniques for integrated and high-performance patient reported adverse drug event extraction. The framework consists of medical entity extraction for recognizing patient discussions of drug and events, adverse drug event extraction with shortest dependency path kernel based statistical learning method and semantic filtering with information from medical knowledge bases, and report source classification to tease out noise. To evaluate the proposed framework, a series of experiments were conducted on a test bed encompassing about postings from major diabetes and heart disease forums in the United States. The results reveal that each component of the framework significantly contributes to its overall effectiveness. Our framework significantly outperforms prior work.

Published by Elsevier Inc.

1. Introduction

In recent years, a growing number of patients are sharing their experiences of healthcare on the Internet. This body of information is described as "cloud of patient experience". The increasing availability of patients' accounts of their care on blogs, social networks, and forums presents an intriguing opportunity to advance the patient-centered care agenda [1]. Patients with chronic diseases such as hypertension, heart diseases, diabetes, and cancer utilize the social media to share their diagnosis, treatment opinions, medications and side effects [2]. Patient self-reports on social media frequently capture medical issues and side effects that clinicians often miss or downgrade. Clinicians' failure to note those issues results in the occurrence of drug non-compliance and preventable adverse events [3]. Mining social media has been considered as a new approach for collecting evidence for drug side effects, drug compliance and drug effectiveness. It can enhance the capture of subjective elements of drug safety and treatment management, providing important insights for clinical practitioners.

The value of patient experience on social media has also drawn attention of researchers from pharmacovigilance community. Pharmacovigilance, also referred to as drug safety surveillance, has been defined as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug problem" [4]. It has predominantly relied on spontaneous reporting systems (SRSs), passive systems composed of reports of suspected ADEs collected from healthcare professionals, consumers, and pharmaceutical companies and maintained by regulatory and health agencies [5]. Two prominent SRSs are the US Food and Drug Administration Adverse Event Reporting System (FAERS) and VigiBase maintained by the World Health Organization (WHO). Other data sources include electronic health records, and publicly available chemical and biological knowledge bases such as DrugBank. Several recent publications attest to the richness of information to be found in patient self-reports of their problems in social media, and also the volume of useful reports is enhanced, thus aiding earlier hypothesis formation and adverse drug event signal detection [6].

Given the clinical and scientific value of patient reports in social media, researchers have begun exploring methods to identify and extract them from social media [7]. Social media contains a large amount of online patient colloquial language. Extracting high quality patient reports of adverse events from such environment can be

* Corresponding author.
E-mail addresses: xiaoliu@email.arizona.edu (X. Liu), hchen@eller.arizona.edu (H. Chen).

http://dx.doi.org/10.1016/j.jbi.2015.10.011
1532-0464/Published by Elsevier Inc.

challenging. Adverse drug events are medical events caused by medications, often presenting as treatment and medical event pairs in patient discussion. They may confound with negated adverse drug events and drug indications. Negated adverse drug events deny causal relation between the drugs and the events. Drug indications are legitimate medical conditions a drug is used for. Discussions of adverse drug events may be based on real patient experience, research, news or hearsay, leading to noise and a significant number of duplicates [8]. Table 1 illustrates these issues with posts from online forum of American Diabetes Association.

From Table 1, we observe that online health consumers adopted their preferred medical terms in forums. These terms are different from medical professional terms (e.g., stroke in post no. 63828 is a consumer preferred term, usually presented as cerebrovascular accident in FAERS; bruising in post no. 34188 is presented as contusion in FAERS). Patient discussions may include different types of drug and event relations. In post no. 63828, the author mentioned stroke and Lipitor. Lipitor is a lipid-lowering agent prescribed to reduce the risk of stroke. Stroke and Lipitor in this post present a drug indication relation. In post no. 9043, patient reported having chest pain when using Actos, presenting an adverse drug event (ADE). Information in forums comes from different sources such as diabetes research (post no. 63828), patient experiences (post nos. 9043, 25139 and 34188), and hearsay (post no. 12200).

Table 1
Examples of patient discussions in social media.

PostID	Post content	Contain ADE?	Report source
9043	I had horrible **chest pain [Event]** under **Actos [Treatment]**	ADE	Patient
12200	From what you have said, it seems that **Lantus [Treatment]** has had some negative side effects related to **depression [Event]** and **mood swings [Event]**	ADE	Hearsay
25139	I never experienced **fatigue [Event]** when using **Zocor [Treatment]**	No	Patient
34188	When taking **Zocor [Treatment]**, I had **headaches [Event]** and **bruising [Event]**	ADE	Patient
63828	Another study of people with multiple risk factors for **stroke [Event]** found that **Lipitor [Treatment]** reduced the risk of stroke [Event] by 26% compared to those taking a placebo, the company said	Drug indication	Diabetes research

Recognizing the importance of mining health social media for pharmacovigilance and current obstacles of extracting patient reported adverse drug events, we are motivated to develop an integrated and high-performance information extraction framework for patient reports of adverse drug effects in health social media. In our proposed framework, we devise a lexicon based medical entity extraction approach, which integrates multiple medical lexicons and consumer health vocabulary for interpreting colloquial health care language. Our major innovation lies in the development of adverse drug event extraction approach using both shortest dependency path kernel based statistical learning method and semantic filtering method with information from medical knowledge bases. This approach, leveraging existing medical knowledge and statistical learning techniques, can significantly increase the precision of extracting adverse drug events. To capture true patient experience, we also develop report source classification to identify actual patient reports of adverse drug events. Our approach identifies patient experienced adverse drug events in social media and provides an efficient way to capture patients' voice in drug safety.

The remainder of this paper is organized as follows. Section 2 introduces a brief research background of prior studies. Section 3 describes our proposed research framework. Section 4 presents evaluation results and discussions. Section 5 concludes this paper.

2. Related work

2.1. Pharmacovigilance in health social media

There has been an increased interest of analyses on health social media content. Leaman et al. [9] explored the value of patient intelligence on pharmacovigilance in social media. Benton et al. [10] acknowledged the demand of advanced techniques for analyzing health social media content. Table 2 summarized the recent work of pharmacovigilance in health social media.

Table 2
Summary of related adverse drug event studies with social media data.

Prior study	Test bed	Focus	Techniques			Results
			Classification	Medical entity extraction	Adverse drug event extraction	
Leaman et al. [9]	Daily strength	AEs	Not applied	Lexicons: UMLS, MedEffect, SIDER	Not applied	Precision: 78.3%; Recall: 69.9%; *F*-measure: 73.9%
Nikfarjam and Gonzalez [11]	Daily strength	AEs	Not applied	Association rule mining	Not applied	Precision: 70%; Recall: 66.3%; *F*-measure: 68.0%
Chee et al. [17]	Health forums in Yahoo! groups	Risky drugs	SVM and Naive Bayes	Lexicons: UMLS, MedEffect, SIDER	Not applied	The ensemble classifier is able to identify risky drugs for FDA scrutiny
Benton et al. [10]	Breast cancer forums	ADEs	Not applied	Lexicons: CHV, FAERS	Co-occurrence based	Promising to detect ADR reported by FDA
Hadzi-Puric and Grmusa [20]	Parenting website	ADEs	Not applied	Lexicons: UMLS	Co-occurrence based	Precision: 75.3%; Recall: 64.7%; *F*-measure: 69.599%
Bian et al. [14]	Twitter	ADEs	SVM	Lexicon: FAERS	Not applied	Accuracy: 74%; AUC value: 0.82
Wu et al. [18]	Online discussions	ADEs	Rocchio method	Lexicon constructed by authors	Generative Model	Extracted ADEs compared to FAERS: precision: 70%; recall: 69%
Mao et al. [2]	Breast cancer forums	ADEs, drug switching	Not applied	Lexicons: CHV, FAERS	Co-occurrence based	Online discussions of breast cancer drugs can help to understand drug switching and discontinuation behaviors
Sarker and Gonzalez [15]	Clinical reports, Twitter, and daily strength	ADEs	SVM	Lexicons:UMLS, WordNet, MedEffect, SIDER, COSTART	Not applied	Achieved detection of sentences with ADE mentions with *F*-measure: 0.812
Segura-Bedmar et al. [16]	Social media	ADEs	Not applied	Lexicon: GATE pipeline	Distant supervision with shallow linguistic kernel	Precision: 48%; Recall: 59%