Artificial Intelligence in Medicine 70 (2016) 62–76

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

An ensemble method for extracting adverse drug events from social media

Jing Liu*, Songzheng Zhao, Xiaodi Zhang

School of Management, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, PR China

ARTICLE INFO

Article history:
Received 7 October 2015
Received in revised form 20 May 2016
Accepted 27 May 2016

Keywords:
Relation extraction
Feature-based approach
Feature selection
Kernel-based approaches
Social media
Adverse drug event extraction

ABSTRACT

Objective: Because adverse drug events (ADEs) are a serious health problem and a leading cause of death, it is of vital importance to identify them correctly and in a timely manner. With the development of Web 2.0, social media has become a large data source for information on ADEs. The objective of this study is to develop a relation extraction system that uses natural language processing techniques to effectively distinguish between ADEs and non-ADEs in informal text on social media.
Methods and materials: We develop a feature-based approach that utilizes various lexical, syntactic, and semantic features. Information-gain-based feature selection is performed to address high-dimensional features. Then, we evaluate the effectiveness of four well-known kernel-based approaches (i.e., subset tree kernel, tree kernel, shortest dependency path kernel, and all-paths graph kernel) and several ensembles that are generated by adopting different combination methods (i.e., majority voting, weighted averaging, and stacked generalization). All of the approaches are tested using three data sets: two health-related discussion forums and one general social networking site (i.e., Twitter).
Results: When investigating the contribution of each feature subset, the feature-based approach attains the best area under the receiver operating characteristics curve (AUC) values, which are 78.6%, 72.2%, and 79.2% on the three data sets. When individual methods are used, we attain the best AUC values of 82.1%, 73.2%, and 77.0% using the subset tree kernel, shortest dependency path kernel, and feature-based approach on the three data sets, respectively. When using classifier ensembles, we achieve the best AUC values of 84.5%, 77.3%, and 84.5% on the three data sets, outperforming the baselines.
Conclusions: Our experimental results indicate that ADE extraction from social media can benefit from feature selection. With respect to the effectiveness of different feature subsets, lexical features and semantic features can enhance the ADE extraction capability. Kernel-based approaches, which can stay away from the feature sparsity issue, are qualified to address the ADE extraction problem. Combining different individual classifiers using suitable combination methods can further enhance the ADE extraction effectiveness.

1. Introduction

An *adverse drug event* (ADE) is defined as any injury due to a medication [1]. This injury can be caused by a medication error, an off-label usage of a medication, or a recommended usage of a medication as per its prescription or label [2]. An adverse drug reaction (ADR), a subtype of an ADE, refers to an unintended response to a drug when it is used at recommended dosage [3]. ADEs are a crucial public health concern, and they could result in increased hospitalizations, morbidity, and even mortality [4,5]. For example, it is estimated that ADEs affect approximately 2 million inpatients each year in the United States [6] and can lead to prolonged hospital stays. In terms of outpatients, ADEs are annually responsible for approximately 125,000 hospital admissions, 1 million emergency visits, and over 3.5 million physician visits in the United States [7]. ADEs can also result in reputation damage for pharmaceutical companies and major financial losses for countries (e.g., approximately $660 million in Australia per year due to estimated 190,000 medication-related hospital admissions [8]). Therefore, detecting ADEs accurately and in a timely manner is important for stakeholders (e.g., patients, physicians, pharmaceutical companies, and regulatory authorities).

Although ADEs can be identified by pre-marketing clinical trials, such trials are limited because they are on selected populations and have constraints on the scale and time [4,9]. Therefore, major

* Corresponding author at: School of Management, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an, Shaanxi 710072, PR China.
E-mail address: liujing2968@163.com (J. Liu).

http://dx.doi.org/10.1016/j.artmed.2016.05.004
0933-3657/

risks with regard to drug safety could remain when the drug hits the market, making post-marketing surveillance the paramount avenue for detecting ADEs that are associated with a drug [10]. Currently, in the United States, post-marketing drug safety monitoring relies primarily on the US Food and Drug Administration (FDA) adverse event reporting system (FAERS), which is a passive system that is populated by voluntary ADE reports from healthcare professionals, pharmaceutical manufacturers, and patients. However, studies have shown that FAERS significantly underestimates the number of ADE cases [11,12] and is incapable of detecting ADEs in a timely manner [12].

In recent years, social media has been an under-explored data source for extracting ADEs. An increasing number of patients (e.g., one-fourth of people with chronic diseases [13]) are turning to social media to seek information, obtain advice, voice concerns, and share experiences concerning drugs [14].

The objective of this paper is to develop a system for extracting ADEs from user-generated content (UGC) on social media using advanced natural language processing (NLP) techniques and machine learning algorithms. Such a system could augment the current passive systems (e.g., FAERS), thereby aiding regulatory authorities in drug-safety related decision making (such as drug recalls, market withdrawals, and safety alerts) as well as reducing legal and monetary loss risks for pharmaceutical companies. Specifically, this study aims to utilize relation extraction methods, which can recognize relationships between two entities in unstructured text, to classify ADE relationship between identified drug entities and event entities from other relations with good performance. We develop a feature-based method that leverages various lexical, syntactic, and semantic features. We also explore the effectiveness of the feature-based method, four well-known kernel-based approaches, and combinations of these methods. Extensive experiments are performed using two health-related discussion forums and one general social networking site (i.e., Twitter) to investigate the feasibility of our proposed system.

2. Related work

2.1. Extracting adverse drug events from social media

ADE extraction research has utilized various data sources, such as spontaneous reporting systems [15], clinical notes [16], and electronic health records [4,17]. However, recent ADE extraction studies are paying attention to social media, such as microblogs (e.g., Twitter [10,11,18–21]) and health-related discussion forums, such as DailyStrength [9,10,21,22], MedHelp [5,23,24], AskaPatient [20,25,26], and American Diabetes Association [26,27].

To conduct automatic ADE extraction from social media, Jiang and Zheng [18] adopted a lexicon-based approach to extract the event entities. However, this study failed to distinguish between ADEs and other types of events (e.g., drug indications). To fill in this research gap, several studies classified recognized events and endeavored to separate ADEs from others. For example, Leaman et al. [9] implemented a rule-based filtering method to remove other relations. Nikfarjam and Gonzalez [22] generated frequent language patterns for expressing ADEs using association rule mining. Bian et al. [11], Sarker and Gonzalez [10], and Nikfarjam et al. [21] explored different deep linguistic features, such as textual features, semantic features, sentiment-related features, and the embedding cluster number feature. Nevertheless, these studies were not able to automatically specify the drug that the identified ADEs were associated with. To address this problem, multiple studies formulated *ADE extraction* as a relation extraction task to detect the specific relationship between the drug entities and event entities, to indicate that "the event is caused by the drug". Co-occurrence, a relation extraction method that is easy to implement, has dominated prior studies [5,19,28–30]. However, co-occurrence generally suffered from low precision. Therefore, recent studies turned to more sophisticated relation extraction approaches. For example, Liu et al. [23,27] adopted the shortest dependency path kernel [31] and achieved an *f*-measure of 66.9%.

2.2. Relation extraction methods

Relation extraction is aimed at recognizing relationships between two entities in unstructured text and has gained considerable attention in the biomedical domain, such as for protein–protein interaction (PPI) extraction [32,33] and drug–drug interaction (DDI) detection [34,35]. Relation extraction has been conducted on various corpora that are derived from the biomedical literature or from newspapers.

There are three main categories of relation extraction approaches: co-occurrence, rule-based, and statistical learning methods [36]. The *co-occurrence approach* is prone to low precision because it assumes that two entities are somehow related if they are mentioned simultaneously [37]. *Rule-based approaches* are prone to low recall because the generated rules could fail to correctly recognize the instances that are expressed in uncovered patterns [38,39]. *Statistical learning methods*, which include feature-based methods and kernel-based methods, generally recast relation extraction as a classification problem and have achieved great success. *Feature-based approaches* leverage various lexical, syntactic, and semantic features, and thus, time- and effort-consuming feature engineering is required [40,41]. *Kernel-based methods*, on the other hand, can address high (or even infinite) dimensional features implicitly by directly computing the inner dot product of the compared instances [42] via a kernel function.

Multiple effective kernels that explore different feature spaces have been proposed. Tree kernel [43], subset tree kernel [44], shortest dependency path kernel [31], and all-paths graph kernel [45] have leveraged shallow parse tree information, syntax tree representation, shortest path connecting two entities in a dependency structure, and both dependency structure as well as linear surface information, respectively. Derivatives of these approaches were developed to overcome the limitations of the original kernels. For example, cosine kernel [46], edit distance kernel [46], walk kernel [47], walk-weighted kernel [48], and dependency trigram kernel [41] have been proposed to overcome the constraint of the shortest dependency path kernel, i.e., the same length requirement on the two compared shortest dependency paths.

Each kernel utilizes a different portion of a sentence's structure [33] and has its own benefits and drawbacks [32]. To take advantage of each kernel and retrieve various important types of information, prior studies have proposed multiple hybrid methods to combine different individual methods, and they indicated that relation extraction capabilities could be enhanced in most cases [32–36,40].

2.3. Research gaps and questions

Several prior studies have formulated ADE extraction from social media as a relation extraction problem and can output "drug/adverse-event" pairs. Although statistical learning methods have achieved great success in the healthcare domain (e.g., PPI extraction and DDI detection), to the best of our knowledge, few studies have utilized these methods to extract ADEs from social media [23,27]. Social media is a very challenging data mining environment; it is abundant in misspellings, colloquial terms, abbreviations, and novel/creative phrases. These intrinsic characteristics of UGC on social media can result in high-dimensional feature space. However, to the best of our knowledge,