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%

Prior studies employed data sources from three categories of social media. Most studies utilized general health discussion forums [9,11]. Others developed research test beds based on disease-focused discussion forums [10,12,2,13]. Benton et al. [10] adopted three breast cancer forums as test bed. Mao et al. [2] developed their test bed on 12 breast cancer forums to understand patient reported adverse drug events. Besides, tweets (microblogs of 140 or fewer characters) have been employed in a recent study [14,15]. Patients sometimes indicate their medications and associated side effects in tweets, presenting real time information for pharmacovigilance. Sarker and Gonzalez [15] developed a text classification method for adverse drug reaction detection on clinical reports, tweets and general health forums. Segura-Bedmar et al. [16] developed a distant supervision approach to extract adverse drug events in Spanish.

Natural language processing techniques adopted in prior studies include text classification, medical entity recognition and adverse drug event relation extraction. For text classification, support vector machines (SVM) and nave Bayes are most commonly used in recent studies. Chee et al. [17] developed ensemble classifiers with SVM and nave Bayes to classify risky drugs and safe drugs based upon online discussions. Bian et al. [14] utilized SVM to filter noise in tweets. Wu et al. [18] developed a discriminant classifier with Ricchio method and a generative model to determine whether the side effect is relevant to the drug.

Medical entity recognition aims to identify medical entities such as treatments and medical problems. Most of the prior studies adopted lexicon-based entity recognition approaches because of the well developed medical lexicons and knowledge bases available in the healthcare domain. The Unified Medical Language System (UMLS) [19] has been adopted in prior studies [9,17,12,20]. Spontaneous reporting systems are often employed to extract treatments and adverse events from text. Medical terms in FDA's Adverse Event Reporting System (FAERS) are used to map drug and adverse event entities in health social media [10,14,2]. MedEffect (Adverse drug event reporting system in Canada) were used to extract adverse events in social media [9,17]. COSTART, a vocabulary created from FAERS, is also adopted in prior study [15]. In health social media research, it is often observed that consumer health vocabularies are different from those of medical professionals [17]. To interpret medical terms in online patient discussions, Consumer Health Vocabulary (CHV), a lexicon linking UMLS standard medical terms to patients' colloquial language [21], is adopted in recent studies [10,22]. Nikfarjam and Gonzalez [11] developed a machine learning based association rule mining algorithm to generate patterns for adverse event recognition.

Adverse drug event extraction utilizes relation extraction techniques to determine if there is a relation between the drug and events and the type of relation is (e.g., drug indications or adverse drug events). Most prior studies have adopted co-occurrence analysis approaches to extract adverse drug event relations [10,22,2]. Benton et al. [10] assumed that if two entities co-occurred within 20 tokens, there was an underlying relation between them.

Most studies evaluated their performance using precision, recall, and *F*1 metrics. For text classification, Bian et al. [14] achieved 74% accuracy in identifying tweets with adverse events. Leaman et al. [9] achieved the best performance values on extracting adverse events from forums with a precision of 78.3%, recall of 69.9% and *f*-measure of 73.9%. Nikfarjam and Gonzalez [11] obtain 70% in precision, 66.32% in recall and 67.96% in *f*-measure using a machine-learning approach to extract adverse events. Sarker and Gonzalez [15] achieved 81.2% *F*-score in identifying sentences with adverse events from text.

To demonstrate the value of adverse drug event reports from social media, researchers have conducted multiple analyses on the extracted results. Benton et al. [10] compared the adverse events extracted by their system against the documented adverse events. Social media ADEs achieves 35.1% in precision, 77% in recall and 52.8% in *f*-measure comparing to the documented ADEs. Chee et al. [17] found patient drug reviews can be used to identify risky drugs on the market and most of the risky drugs they identified are on FDA's drug safety watch list. Yang et al. [22] considered health social media a promising data source for adverse drug event signal detection. Mao et al. [2] found online discussions of breast cancer drugs can help to understand drug switching and discontinuation behaviors.

Based on our review of prior research, we find machine learning based classification techniques are widely adopted in social media research to filter out noise. Medical entity extraction with medical lexicons and ontologies achieves satisfying performance. Co-occurrence analysis-based approach for extracting adverse drug events has clear limitations. This approach captures little syntactic or semantic information. As a result, it can generate false adverse drug events when negation exists in sentences. The extracted adverse drug events can be confounded with drug indications. This approach is not able to precisely capture adverse drug events when multiple adverse event entities appear in the same sentence. Although there are duplicated reports caused by news and third-hand accounts in health social media, none of the prior studies addressed this issue.

Our analysis of these studies motivated us to incorporate the following components in our proposed framework: the development and evaluation of a scalable and semantic-rich relation extraction method for adverse drug event extraction and a robust report source classification method to identify adverse drug events based on actual patient experience.

2.2. Biomedical relation extraction

Automatically extracting biomedical information has been the subject of significant research efforts due to the rapid growth in biomedical development and discovery [23]. Biomedical relation extraction techniques are often developed to identify relations such as gene-disease relations and protein interactions from text.

Biomedical relation extraction techniques are often categorized into four types: co-occurrence analysis, rule-based approaches, statistical learning approach and hybrid approach, which utilizes both rules and statistical learning. For each type, methods vary in how they utilize the lexical, syntactic, and semantic information in text. Co-occurrence analysis identifies relations between biomedical entities based on their probability of occurrence in text. This approach assumes that if two entities are both mentioned within a certain range there is an underlying biological relationship [2]. In most cases, only lexical information is needed for co-occurrence analysis. Due to their simplicity and flexibility, these approaches have been widely used for relation extraction and can achieve high recall. Since it uses little syntactic information or semantic information, co-occurrence analysis often achieves low precision. Co-occurrence analysis approach has been used to identify adverse drug events in health social media [22,2].

In rule-based approaches, researchers have manually developed rules based on syntactic or semantic information to parse relations. Syntactic parsing approaches extensively utilize syntactic rules for relation extraction [24,25]. Another rule-based method relies on semantic information in sentences. Semantic indicators of biomedical relations consist of certain slots of trigger words (e.g., 'interact with' or 'bind to') manually developed by experts. A pair of entities which satisfies a certain predefined template is identified as a relation. In biomedical domain, semantic parsers for relation extraction are often applied because of the availability of semantic information in biomedical literature and knowledge bases [26].

Statistical learning approach view relation extraction as a text classification problem and requires little or no manual development of rules or templates. Patterns are learned from a corpus of documents in which human experts have tagged the desired relations. Statistical learning can be categorized into feature-based methods and kernel-based methods.

For feature-based methods, each relation instance is represented as a feature vector $X = \{x_1, x_2, \ldots, x_n\}$ in an n-dimensional space. Features are defined and selected to capture the data characteristics. Yang et al. [27] used bag-of-words features, medical related words, and entity distance as features. Bui et al. [24] adopted bag-of-words features, part-of-speech tag features and entity distance features to classify relation types.

Kernel-based methods are an effective alternative to explicit feature extraction. They retain the original representation of relation instances and use the object only via computing a kernel function between a pair of instances [28–32]. Tree kernels leverage the dependency tree of the entire sentences [32] while shortest dependency path kernels only concern with the minimal proportion that bridges two entities [33]. They assume that the most important features for relation extraction concentrated on the shortest path between two entities on the dependency tree. Li et al. [31] developed a composite kernel that combines linear, sequence, and tree kernel for gene-disease relation extraction. Miwa et al. [30] used composite kernel formed by Bag-of-words kernel, sub tree kernel, shortest dependency path kernel and graph kernel to extract protein interactions. Thomas et al. [29] applied ensemble learning on graph kernel, shortest dependency path kernel and shallow linguistic kernel to extract drug–drug interaction from medical literature.

Among various biomedical relation extractors, both rule-based approach with syntactic or semantic information and statistical learning approach have shown good performance. Statistical learning can automatically learn relation patterns from annotated corpora. Kernel-based learning methods have shown promise in identifying various biomedical relations such as protein interactions and gene-disease relations. They achieve better performance than feature based approached as they utilized the syntactic and semantic information, which can concisely and precisely capture the relationships between entities.

In social media, online users utilize a large amount of colloquial language, which can lead to a large sparse lexical feature set and low performance for feature based approaches. However, these discussions still follow certain syntactic and semantic patterns. We believe that kernel based learning method can be used for extracting adverse drug event from noisy social media text with help of the syntactic and semantic representation of the data. When processing health social media text, drug indications are often confounded with adverse events in adverse drug event extraction. The difference between drug indications and adverse events cannot be captured by syntactic and semantic parsing but domain knowledge bases. Drug regulatory agencies have the indications of marketed medications documented in medical knowledge bases. Adding semantic filtering with the drug indication information from medical knowledge bases may further enhance the performance of statistical learning based adverse drug event extraction.

2.3. Text classification

A common challenge for social media research is to extract high quality information from noisy data, especially for health social media research. Many prior studies in health social media domain utilized text classification techniques to differentiate information sources and extract high quality proportion. To explore how text classification can help to identify patient ADE reports in social media, we review the most relevant recent studies using text classification techniques in health social media research.

With respects to the test beds used, classification technique has been applied to health social media data such as Yahoo Answers [34], micro blogs from Twitter [14], and online health communities [17,35]. Text classification techniques have been applied in prior studies for a variety of objectives. Automatic classifications have been developed to classify health consumer posts and health professional posts in Yahoo Answers [34], to identify whether the author is using certain drugs in tweets [14]. It also has been used to classify threads from patient community members or moderators in online health community [35].

The most commonly used features in text classification are bag-of-words [14,35,34]. For unique test beds such as Twitter, hashtags and URLs are adopted in the feature set. Huh et al. [35] utilized sentiment related words from LIWC as features in classification. In terms of learning methods, the most commonly employed learning algorithms are Support Vector Machines (SVM) [14,17,34], nave Bayes [17,35]. Liu et al. [34] achieved the best performance with an f-measure of 89.1% when classifying Yahoo Answers posts into patients' posts or medical professionals' posts. Bian et al. [14] gained 82.0% in f-measure for identifying drug user from tweets. Based on our review of prior studies, text classification techniques can effectively identify health consumer posts from health professional posts in Yahoo Answers and identify drug users from tweets, which are close to the task of identifying adverse drug events based on patient experience. Bag-of-words is the most commonly used and effective features in representing the instances for classification.

2.4. Research gaps and questions

Based on our review, we have identified several research gaps. First, little advanced statistical learning based relation extraction has been adopted in health social media adverse drug event research. Prior pharmacovigilance research in health social media employed co-occurrence analysis based relation extraction techniques to extract adverse drug events. Co-occurrence analysis only utilizes lexical features in the sentences. Neglecting the syntactical information and semantic information in the sentence, this approach could generate significant false positive results causing by negations and true drug indications. Second, the adverse drug events discussed in health social media may come from a variety of sources such as patients, news, research, and stories from third-hand accounts, which result in redundancy and noise. While more researchers started to be aware of the importance of patients' voice in drug safety reporting [3], few prior studies in social media pharmacovigilance research identified patient adverse drug event reports based on true patient experiences. The value of health social media, an open and popular platform for patient to speak out their problems and demands, has not been fully explored.

Based on the research gaps identified, we proposed the following research questions:

1. How can we develop an integrated and scalable research framework for mining patient reported adverse drug events from patient forums?
2. How can statistical learning techniques augmented with health-relevant semantic filtering improve the extraction of adverse drug events as compared to other baseline methods?
3. How can we identify true patient reported adverse drug events among noisy forum discussions?

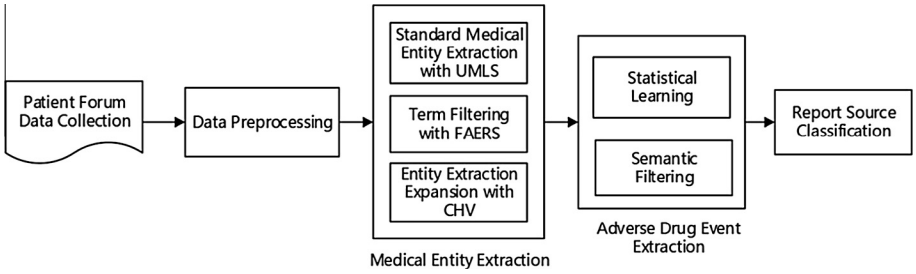

Fig. 1. Research framework for pharmacovigilance in health social media.

3. Research method

Our proposed research framework for pharmacovigilance in health social media is illustrated in Fig. 1. Major components are explained in detail below.

3.1. Patient forum data collection

We developed an automated crawler program to download web pages from online patient forums. An extractor program was written to extract specific fields in patient discussions. Collected information includes post ID (the unique identifier of a post in the forum), URL, topic title, post author's ID (the unique identifier of a user in the forum), post date, and post content [36].

3.2. Data preprocessing

Data preprocessing prepares the raw data for subsequent analysis. Data preprocessing consists of two steps: text cleaning and sentence boundary detection. We developed regular expression base approach to remove URLs, duplicate punctuations, and personally identifiable information such as email addresses, social security numbers, and phone numbers from the text. As our study focuses on sentence level information extraction and processing, we segment each post into sentences with a natural language processing toolkit, OpenNLP.[1] OpenNLP provides state-of-the-art machine learning based sentence boundary detection algorithm. Each post we obtain from the crawler can then be segmented into individual sentences with OpenNLP.

3.3. Medical entity extraction

It is a challenging task to extract medical entities from noisy patient-generated content. Leaman et al.'s [9] lexicon based approach was the best performing medical entity recognition system in prior studies. We apply multiple types of lexicon sources to extract drug names and adverse events from the text, including UMLS [19], FAERS,[2] and CHV.[3]

MetaMap,[4] a Java API from the National Library of Medicine, is used to identify medical concepts in UMLS from health social media. Currently, the UMLS has 135 semantic types,[5] which are further abstracted into 15 semantic groups,[6] such as 'Chemicals and Drugs,' 'Disorders,' and 'Genes & Molecular Sequences'. We configure MetaMap to recognize the terms that belong to the 'Chemicals and Drugs' semantic group for drug name entities and 'Disorders' group for adverse event entities. We start with MetaMap to identify medical entities matching standard medical lexicons in patient forums.

Results of MetaMap to extract terms belong to semantic group 'Chemicals and Drugs' and 'Disorders' contain some false positive information. Food and recipe ingredients in the forum discussions are often identified as 'Chemicals and Drugs'. Common verbs such as 'find' and 'have' can be extracted as 'Disorders'. To avoid these issues, we filter results from MetaMap with drug names and event names from the FDA's drug safety database, FAERS. Medical entities that never appear in FAERS are removed from further analysis.

Patient discussions of medical problems are different from those in medical documents. They contain consumer-preferred description of medical terms. To understand patient discussions in social media, we incorporate the Consumer Health Vocabulary (CHV), which contains 47,505 UMLS standard medical terms, corresponding to 127,081 consumer-preferred terms. For each medical entity remains, we query the CHV to get its consumer-preferred terms, which cannot be recognized by MetaMap. We then look up the forum collection with these consumer-preferred terms to extend our medical entity extraction to incorporate consumer-preferred way of mentioning medical terms. Later, we normalize those mentions in consumer-preferred terms to standard medical terms. All sentences with both drug and event entities are extracted for further analysis.

3.4. Adverse drug event extraction

Patients' adverse drug event discussions in forums are more informal and colloquial than biomedical literature or clinical notes, which require medical knowledge and complex linguistic techniques to interpret. Based on our review of prior biomedical relation extraction studies, a hybrid approach with both statistical machine learning methods and rule-based filtering achieved satisfying performance [24]. Our approach incorporates the statistical learning method for relation detection and semantic information from medical and linguistic knowledge bases to identify adverse drug events from drug indications and negated ADEs. The statistical learning and semantic filtering components for adverse drug event extraction are presented in Fig. 2.

3.4.1. Statistical learning

Statistical learning of adverse drug event extraction determines whether a drug and a medical event in one sentence have a relation. We developed a shortest dependency path kernel based statistical learning method in our framework. Shortest dependency path kernel function has shown promise in identifying various relations such as gene interactions and drug interactions in prior studies [29,30,33]. Yet it hasn't been used in extracting adverse drug events. We utilized Support Vector Machines (SVM) to learn patterns from posts with related drugs and events. Statistical learning component includes feature generation, kernel function, and classification method.

[1] https://opennlp.apache.org/.
[2] http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/.
[3] http://www.consumerhealthvocab.org/.
[4] http://metamap.nlm.nih.gov/.
[5] http://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt.
[6] http://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt.

Feature generation. In health social media, patients mentioned their adverse drug events in colloquial languages, which make lexical and distance features less effective in statistical learning due to the data sparsity. However, patients' narratives about adverse drug events still follow certain syntactic and semantic rules. We propose to extract syntactic and semantic features from sentence dependency parse trees to represent the instances. Dependency parsing generates word-to-word links based on grammatical relations. They represent both syntactic and semantic information between words in a sentence. In dependency parse trees, the syntactic dependency shows in the hierarchical structures of the trees. The semantic dependency is demonstrated by the directions of the links. In this study, Stanford Parser[7] was used for dependency parsing. A grammatical relation holds from a dependent to a governor (also known as a regent or a head). Fig. 3 shows the dependency tree of a sentence. In this sentence, nausea is an adverse event entity and Byetta is a diabetes treatment. Grammatical relations between words are illustrated in the figure. For example, 'nausea' is the direct object of 'gotten,' thus they have a grammatical relation 'dobj.' In this case, 'gotten' is the governor and 'nausea' is the dependent.

Algorithm 1. Shortest Dependency Path Extraction

Inputs:
  A relation instance i, a pair of related drug and event $R(drug, event) = True$, and dependency graph T
Outputs:
  $Path$, the shortest dependency path from event to drug
procedure ShortestDependencyPathExtraction(i, $drug$, $event$, T)
  if $drug \in event$.dependents() **then**
    $Path \leftarrow \{event, \leftarrow, drug\}$
    return $Path$
  else
    $Path \leftarrow \{event\}$
    $End \leftarrow \{drug\}$
    $Head \leftarrow \{event\}$
    $Tail \leftarrow \{drug\}$
    while $Head \neq Tail$.governor **do**
      if $drug \in Head$.dependents() **then**
        $Head \leftarrow Head$.governor
        $Path \leftarrow Path + \{\rightarrow, Head, \leftarrow, drug\}$
        return $Path$
      else
        $Head \leftarrow Head$.governor
        $Path \leftarrow Path + \{\rightarrow, Head\}$
      if $event \in Tail$.governor.dependents() **then**
        $Tail \leftarrow Tail$.governor
        $End \leftarrow \{event, \rightarrow, Tail, \leftarrow\}$ + End
        return $Path$
      else
        $Tail \leftarrow Tail$.governor
        $End \leftarrow \{Tail, \leftarrow\}$ + End
        $Path \leftarrow Path + \{\leftarrow\} + End$
        return $Path$

A large proportion the dependency tree is not relevant to the relation of medication and medical condition in the sentence. Prior studies in relation extraction show that the contribution of dependency tree in establishing the relationship between two entities is almost exclusively concentrated in the shortest path between them on the dependency tree [33,29]. To utilize the shortest path between medical event entity and drug entity in the dependency

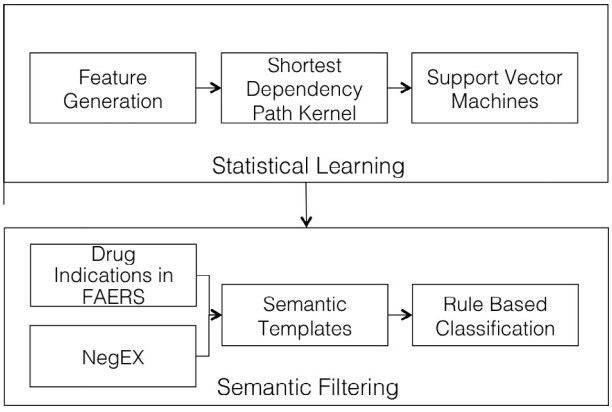

Fig. 2. Procedures for adverse drug event extraction.

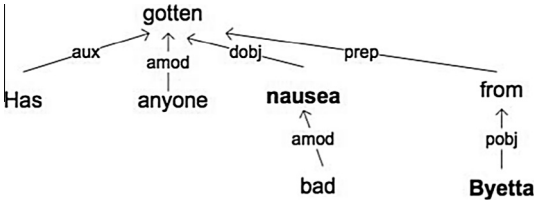

Fig. 3. A sample sentence represented as a dependency tree.

tree (shortest dependency path), we propose Algorithm 1 to extract the shortest paths of two entities from dependency trees. It searches for the shortest path from medical events to treatments on the dependency tree for each relation instance and captures not only the words but also dependency directions on the path.

Syntactic and semantic class mapping. To increase the robustness of our method, we expanded shortest dependency path by categorizing words on the path into word classes with varying degrees of generality. Word classes include part-of-speech (POS) tags and generalized POS tags. POS tags are extracted with Stanford CoreNLP packages.[8] We generalized the POS tags with Penn Tree Bank guidelines for the POS tags. Semantic types (Event and Treatments) are also used for the two ends of the shortest path. Table 3 lists all the POS tags and generalized POS tags from our data set.

The feature representation of relation instances can be defined as the Cartesian product of all the elements on the path. The feature representation of the sample sentence in Fig. 3 is illustrated in Eq. (1). The original sentence thus can be represented in a sequence as $X = [x_1, x_2, x_3, x_4, x_5]$, where $x_1 = \{Nausea, NN, Noun, Event\}$, $x_2 = \{\rightarrow\}$, $x_3 = \{gotten, VBD, Verb\}$, $x_4 = \{\leftarrow\}$, $x_5 = \{Byetta, NN, Noun, Treatment\}$.

$$\begin{bmatrix} Nausea \\ NN \\ Noun \\ Event \end{bmatrix} \times [\rightarrow] \times \begin{bmatrix} gotten \\ VBD \\ Verb \end{bmatrix} \times [\leftarrow] \times \begin{bmatrix} Byetta \\ NN \\ Noun \\ Treatment \end{bmatrix} \quad (1)$$

Shortest dependency path kernel function. Statistical learning methods rely on kernel functions to find a hyperplane that separates positive instances from negative. For shortest dependency path kernels, if $x = x_1x_2x_3x_4 \cdots x_m$ and $y = y_1y_2y_3y_4 \cdots y_n$ are two relation instances, where x_i denotes the set of features corresponding to position i, the kernel function is defined as in Eq. (2):

$$K(x, y) = \begin{cases} 0 & m \neq n \\ \prod_{i=1}^{n} C(x_i, y_i) & m = n \end{cases} \quad (2)$$

[7] http://nlp.stanford.edu/software/lex-parser.shtml.

[8] http://nlp.stanford.edu/software/corenlp.shtml.

Table 3
POS tags and generalized POS tags.

Part-of-Speech (POS) tags	Generalized POS tags
CC	Conjunction
CD	Number
DT, PDT	Determiner
IN	Preposition
JJ, JJR, JJS	Adjective
NN, NNS, NNP, NNPS	Noun
POS	Possessive ending
PRP, PRP$	Pronoun
RB, RBR, RBS	Adverb
RP	Particle
TO	to
UH	Interjection
VB, VBD, VBG, VBN, VBZ, VBP	Verb
WDT, WP, WP$, WRB	Wh-words
EX, FW, LS, MD, SYM	Others

$C(x_i, y_i) = |x_i \cap y_i|$ is the number of common features between x_i and y_i.

Algorithm 2 details the shortest dependency path kernel function we developed.

Algorithm 2. Shortest Dependency Path Kernel Function

Inputs:
Relation instance $x = x_1x_2x_3 \cdots x_m$ and relation instance $y = y_1y_2y_3 \cdots y_n$
Outputs:
$K(x, y)$, similarity scores between x and y
procedure ShortestDependencPathKernelFunction(x, y)
  if $m \neq n$ **then**
    $K(x, y) \leftarrow 0$
  else
    while $i \leqslant m$ **do**
      $K(x, y) \leftarrow K(x, y) \times |x_i \cap y_i|$
  return $K(x, y)$

For instance, relation instance x = {When this happens, the basal action of your Lantus could cause hypoglycemia.} can be represented as x = [{Hypoglycemia, NN, Noun, Event}, {→}, {cause, VB, Verb}, {←}, {action, NN, Noun}, {←}, {Lantus, NN, Noun, Treatment}]. Relation instance y = {But, now I've read a few posts in this thread that indicate depression as a possible side effect from Lantus.} can be represented as y = [{depression, NN, Noun, Event}, {→}, {indicate, VBP, Verb}, {←}, {effect, NN, Noun}, {←}, {Lantus, NNP, Noun, Treatment}]. $K(x, y)$ can be computed as the product of the number of common features x_i and y_i in position i. $K(x, y) = 3 \times 1 \times 1 \times 1 \times 2 \times 1 \times 3 = 18$. Based on the result, we can see relation instance x and y have a very high similarity score. If relation instance x has a drug event relation, relation instance y is very likely to contain a drug-event relation as well.

Classification. Classification in relation detection aims to distinguish relation instances with a relation from those without any relationship. We adopted Transductive Support Vector Machines (TSVM) [37] for classification in relation detection. SVM-light, an open source software package for Transductive Support Vector Machines is applied in this study because it is widely used in prior studies and enables users to define customized kernel functions [31].

We customized SVM-light by adding our shortest dependency path kernel function. We trained the TSVM classifier on the shortest dependency path kernel and then applied this classifier to identify instances with a drug-event relation. The procedures of statistical learning are summarized in Algorithm 3.

Algorithm 3. Statistical Learning Algorithm

Inputs:
All relation instances I with at least a pair of *drug* and *event*
Outputs:
Whether a pair of *drug* and *event* are related, $R(drug, event) = True$ or *False*
procedure StatisticalLearningAlgorithm$(drug, event)$
  for each pair of drug and event, $R(drug, event)$ **do**
    Generate dependency graph T of instance i containing $R(drug, event)$
    $Path \leftarrow ShortestDependencyPathExtraction\ (i, drug, event, T)$
    $Feature \leftarrow$ Syntactic and Semantic Classes Mapping$(Path)$
  Separate relation instances into training set and test set
  Train a SVM classifier C with shortest dependency kernel function on Training set
  Use the SVM classifier C to classify instances in test set into two classes $R(drug, event) = True$ and $R(drug, event) = False$

3.4.2. Semantic filtering

Shortest dependency path kernels can detect related drug and medical events. However, this method cannot precisely capture negation in sentences and differentiate drug indication relations from adverse drug events. Most prior studies neglected the importance of filtering out drug indications and negated ADEs for analysis, leading to a low precision. To address these issues, we develop a semantic filtering algorithm, which utilizes the semantic knowledge in drug safety database to remove drug indications and rules from negation detection tool to filter out negated ADEs.

As drug indications are regularized and well-documented in drug safety databases such as FAERS, we acquired drug indication knowledge from FAERS to formulate templates and filter drug indications. For negation detection, we utilized the linguistic rule-based negation detection tool, NegEx [38,39]. NegEx is a nature language processing system for negation detection of medical events in discharge summaries. NegEx has been adopted in prior studies for annotating biomedical text [40] and identifying medical events from medical discharge records [41]. The detailed procedures for semantic filtering are presented in Algorithm 4.

Algorithm 4. Semantic Filtering Algorithm

Inputs:
a relation instance i with a pair of related *drug* and *event*, $R(drug, event) = True$
Outputs:
$T(drug, event)$, relation type between *drug* and *event*
procedure SemanticFilteringAlgorithm$(drug, event)$
  if $drug \in FAERS.drug()$ **then**
    $indications \leftarrow FAERS.indication(drug)$
    if $event \in indications$ **then**
      return $T(drug, event)$ = drug indication
    for rule ∈ NegEx **do**
      if instance i matches rule **then**
        return $T(drug, event)$ = negated adverse drug event
  else
    return $T(drug, event)$ = adverse drug event

3.5. Report source classification

To reduce noise and redundancy, report source classification is proposed to filter ADE reports not grounded in patients' experiences. There is no previous health social media research that has addressed this issue. Based on our review of prior studies, text classification techniques can effectively identify health consumer posts from health professional posts in Yahoo Answers and identify drug users from tweets, which is close to the task of identifying adverse drug events based on patient experience [14,13].

In order to classify the report source of adverse drug events, we developed a feature-based classification model to distinguish patient reports from hearsay. We adopted BOW features and Transductive Support Vector Machines for classification. Semi-supervised classification methods such as Transductive SVM, which leverages both labeled and unlabeled data can build the model with a small set of annotated data and conduct transductive inference in unlabeled data [37]. It is more scalable than traditional supervised methods because of the large amount of unlabeled data available in social media.

3.6. Research hypotheses

Based on the research gaps identified in the literature review, we believe our proposed framework can significantly improve the performance of patient adverse drug event extraction in health social media. In particular, we propose the following hypotheses:

H1a. Statistical learning methods in adverse drug event extraction will outperform the co-occurrence analysis based approach.

H1b. Semantic filtering in adverse drug event extraction will further improve the performance of adverse drug event extraction.

H2. Report source classification (RSC) can improve the results of patient adverse drug event report extraction as compared to not accounting for report source issues.

4. Experiment and results

4.1. Research test bed

Chronic diseases, such as diabetes and heart diseases rely on patient self-management. Many online health forums have emerged to provide chronic disease patients with an anonymous connection to an understanding audience where they can ask questions, gain knowledge, and share frustrations about their treatments. Our research test bed is developed from major diabetes patient forums and heart disease discussion boards in the United States, including American Diabetes Association online community, Diabetes Forums, Diabetes Forum, Heart Disease, Heart Rhythm and Coronary Heart Disease discussion boards from MedHelp.

The American Diabetes Association (ADA) is a United States-based association working to fight the consequences of diabetes and to help those affected by diabetes. Online community of ADA attracts over 7000 registered users and thousands of guest visitors. Diabetes Forum is a large online community for diabetes patients with about 25,000 registered users. Diabetes Forums is a diabetes support forum founded in 2002 with over 50,000 registered users. Majority of the users on these sites are diabetes patients and some of them are diabetes caregivers. MedHelp is a general health online community for consumer health information founded in 1994. It contains a wide range of topics from herbal remedies to medication experiences. The content of MedHelp is organized by disease discussion boards. User comments from these boards require noise filtering before further analysis. Disease-focused platforms such as ADA online community, Diabetes Forum, and Diabetes Forums contain more concentrated treatment discussions and disease-relevant patient profiles, presenting great potential for identifying adverse drug events for specific diseases. All of these communities and forums have moderators to ensure the discussion quality and filter spams and advertisements. A summary of test bed is shown in Table 4.

Table 4
Summary of test bed.

Forum name	Number of posts	Number of topics	Time span	Total number of sentences
American Diabetes Association	184,874	26,084	2009.2–2014.1	1,348,364
Diabetes Forums	568,684	45,830	2002.2–2014.1	3,303,084
Diabetes Forum	67,444	6474	2007.2–2014.1	422,355
MedHelp (Heart Diseases)	251,472	66,012	1995.2–2014.1	2,118,101

Table 5
Summary of medical entities in annotated data.

Forum name	Drug	Medical event	Total number of mentions
American Diabetes Association	312	284	596
Diabetes Forums	302	296	598
Diabetes Forum	274	245	519
MedHelp	321	343	664

4.2. Evaluation metrics

We adopt standard machine-learning and text analysis evaluation metrics, precision, recall and *f*-measure, to evaluate the performances of our framework. These metrics have been widely used in information extraction and health social media studies [31,36].

4.3. Experiments

In this study, we conduct our experiments on extracting patient reports of adverse drug events by performing three tasks: medical entity extraction, adverse drug event extraction, and report source classification. We conduct 5-fold cross validation to obtain the evaluation results for adverse drug event extraction and report source classification. For each forum, each time 80% of labeled data and all the unlabeled sentences in our test bed are used as training set and 20% of labeled data are used as test set.

4.3.1. Medical entity extraction

To evaluate the performance medication entity extraction, we selected 250 sentences with at least one drug name from each forum. They are annotated for medical entities as a gold standard. A research associate in pharmacy established definitions and content coding for labeling entities and medical event entities.[9] Two graduate level research associates were trained to annotate the selected sentences for medical entities. When their labels disagreed, a third rater would review the data and make a final decision. A summary of the annotated data is provided in Table 5.

4.3.2. Adverse drug event extraction

To conduct relation detection, we randomly selected 400 sentences with at least one drug entity and one medical event entity from each forum for annotation. With this approach, we focus on determine the relation of drugs and medical events in the same

[9] https://github.com/vulix/AZDrugMiner.

sentences. Relations of drugs and medical events across sentences in the same post are not considered in this study.

Content coding for labeling these sentences is established based on information in existing knowledge bases and advice from clinical experts. Each pair of drug and medical event in the sentences is considered as a relation instance. Two research associates annotated these relation instances with arbitration by a third rater in cases in which the first two disagreed. A summary of the annotated relation instances is provided in Table 6.

To demonstrate the efficacy of our approach, we conduct co-occurrence analysis-based adverse drug event extraction as a baseline for comparison. We adopted the approach from a prior study, in which if a drug occurred with 20 tokens of an event term, then this was treated as a co-occurrence [10].

4.3.3. Report source classification

For report source classification evaluation, we use the same 400 sentences in prior experiment with at least one drug entity and one medical event entity from each forum as labeled data. We established definitions and decision rules for labeling whether the description in each sentence is based on patients' own experiences or not. Two research associates were trained to label the selected sentences from each forum based on these rules. A summary of the annotated data is in Table 7. Each sentence represents a classification instance.

Table 6
Summary of drug and event relations in annotated data.

Forum name	Has a relation			No relation	Total
	Adverse drug event	Drug indication	Negated ADE		
American Diabetes Association	276	169	15	302	762
Diabetes Forums	245	125	23	257	650
Diabetes Forum	223	139	12	286	660
MedHelp	336	186	5	220	747

Table 7
Summary of report sources in annotated data.

Forum name	Patient report	Others
American Diabetes Association	239	161
Diabetes Forums	224	176
Diabetes Forum	274	126
MedHelp (Heart Diseases)	256	144

Table 8
Results of medical entity extraction.

Forum name	Entity type	Precision (%)	Recall (%)	*F*1 (%)
American Diabetes Association	Drug	93.0	91.7	92.3
	Medical event	87.3	80.3	83.6
Diabetes Forums	Drug	92.5	87.1	89.7
	Medical event	86.5	78.7	82.5
Diabetes Forum	Drug	91.4	86.4	88.8
	Medical event	85.4	76.5	80.7
MedHelp (heart disease)	Drug	94.3	81.3	87.3
	Medical event	79.3	73.5	76.3

4.4. Results and discussions

4.4.1. Medical entity extraction

We compared the results from our automatic tagger against the manual annotation for each forum. Table 8 shows the experiment results on all four forums.

Our approach achieved average 90% in *f*-measure for drug entity extraction and average 80% in *f*-measure for medical event extraction. The performance of our approach was attributable to the incorporation of consumer health vocabulary and knowledge-based filtering with the FAERS drug safety database. Our approach performed better on Diabetes forums than MedHelp heart disease discussion board. Diabetes forums mainly focused on diabetes related treatments and medical events, thus they usually have a high consistency in terminology. MedHelp is a general health social website where users are from diverse background. Thus it has more diverse health vocabulary and higher linguistic creativity, which may cause more errors. The medical events identified by our approach may still have negation issues and drug indications which will be filtered in later stage, thus not strictly comparable to the results in other studies.

Based on our evaluation, errors in drug entity identification mainly occurred due to spelling errors and short names for medications. We can observe that medical entity extraction on event entities attains a lower performance than extraction on drug entities. The major source of error in extracting events is caused by patients' ambiguous descriptions of medical events (e.g., 'hypo symptoms' and 'a low', which stands for hypoglycemia). To further improve the performance, more advanced machine learning based named entity taggers are needed.

4.4.2. Adverse drug event extraction

We compared the baseline co-occurrence method (CO) against statistical learning method (SL) as well as against our proposed adverse drug event extraction method including statistical learning and semantic filtering (SL + SF). Table 9 shows the performance results for the three different methods on extracting adverse drug events.

Based on the evaluation results, we observe that our approach consistently increases the precision and *f*-measure for adverse drug event extraction across four forums. Statistical learning contributes to increased precision while leading to lower recall. Semantic filtering further increases the precision without affecting the recall.

Our proposed approach's *f*-measure is about 10% higher than the co-occurrence analysis approach. The precision of our approach is about 37% higher than the co-occurrence analysis approach. The precision of co-occurrence analysis method is dependent on the data set. It ranges from 38% to about 45% because of the different levels of medical information richness in the discussions. Users

Table 9
Results of adverse drug event extraction.

Forum name	Method	Precision (%)	Recall (%)	*F*1 (%)
American Diabetes Association	CO	36.2	100.0	53.2
	SL	62.0	56.5	59.1
	SL + SF	82.0	56.5	66.9
Diabetes Forums	CO	37.7	100.0	54.8
	SL	64.2	60.4	62.2
	SL + SF	78.6	60.4	62.2
Diabetes Forum	CO	33.8	100.0	50.5
	SL	62.5	58.0	60.2
	SL + SF	75.2	58.0	65.5
MedHelp (heart disease)	CO	44.9	100.0	62.0
	SL	65.4	65.3	65.3
	SL + SF	80.7	65.3	72.2

discussed not only their treatment side effects but the indications as well. Sometimes the discussions may involve a large number of drug names, leading to low precision for co-occurrence analysis approach. For pharmacovigilance research, it may be more meaningful to capture adverse drug events precisely than to get a large amount of false reports. Our approach managed to increase the precision of extraction and improve the quality of extracted health social media adverse drug event reports.

We also observe a decrease in recall (from 100% to about 60%) while incorporating the kernel-based statistical-learning method. The low recall is often caused by errors in detecting relations in long relation instances. Those long relation representations have low occurrences in labeled data, resulting in a low learning rate and a low recall. This issue can be resolved by incorporating active learning, a form of machine learning, which determines what relation instances, should be labeled for better extraction performance.

4.4.3. Report source classification

Performance of report source classification (RSC) on extracting patient self-reports is listed in Table 10. Without report source classification (RSC), the performance of extraction is heavily affected by noise in the discussion. The precision ranged from 56% to 69% without RSC. Overall performance (*F*-measure) ranged from 71.8% to 77.7%. After report source classification, the precision and *F*-measure significantly improved. The overall performance (*F*-measure) increased to above 80%. Errors in report source classification mainly occurred in long sentences with ambiguous description of information source or short sentence with implicit information for who had the event. Error in report source classification may be relieved by using more contextual based semantic analysis across sentences.

Table 10
Results of report source classification.

Forum name	Method	Precision (%)	Recall (%)	*F*1 (%)
American Diabetes Association	With RSC	83.9	84.3	84.1
	Without RSC	59.7	100.0	74.8
Diabetes Forums	With RSC	87.2	83.1	85.1
	Without RSC	56.0	100.0	71.8
Diabetes Forum	With RSC	86.5	86.4	86.4
	Without RSC	68.5	100.0	81.3
MedHelp (heart disease)	With RSC	89.6	91.4	90.5
	Without RSC	63.5	100.0	77.7

Table 11
Significance test on *F*1 for hypotheses testing.

Significance test on *F*1		
	Hypothesis	Significance level
H1a	SL > CO	0.027*
H1b	SL + SF > SL	0.021*
H2	RSC > Without RSC	0.011*

4.5. Hypotheses testing

In order to test our hypotheses, we conducted pair-wise one tailed *t*-tests on *F*-measure. Bootstrapping is used for statistical testing. Bootstrapping was performed by randomly selecting 40 instances for testing and remaining 360 for training, 50 times. We evaluated the *F*-measure using pair wise *t*-tests across the samples ($n = 50$). The p values for the tests of our hypotheses for adverse drug event extraction and report source classification are presented in Table 11.

For H1a, statistical learning method (SL) in adverse drug event extraction outperforms the baseline co-occurrence analysis approach (CO). The p value *F*-measure < 0.05. The improvement in overall performance is significant. H1a is supported. For H1b, semantic filtering in adverse drug event extraction (SL + SF) further improves the performance of statistical learning based (SL) adverse drug event extraction. The p value for *F*-measure is less than 0.05. H1b is supported. For H2, report source classification (RSC) can improve the results of patient adverse drug event report extraction as compared to not accounting for report source issues. The p value for *F*-measure is significant. H2 is supported.

4.6. Comparing our proposed framework to prior co-occurrence based approach

To evaluate the impact of our approach on patient reported ADE extraction, we analyze the results of our framework. Fig. 4 shows the changes in number of reports when applying our framework across four forums in our dataset.

There are a large number of false adverse drug events which couldn't be filtered out by co-occurrence based approach. Based on our approach, only 33–40% of all the relation instances contain adverse drug events. Among them, about 50% comes from patient reports. Based on our analysis on the result, we observe that our research framework is very effectively in extracting patient ADE reports in social media. It has significantly reduced the noisy

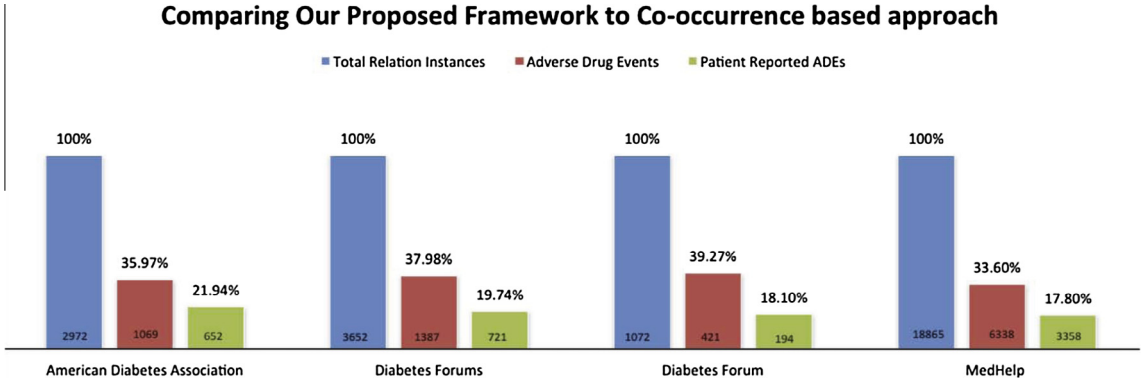

Fig. 4. Comparing our proposed framework to prior co-occurrence based approach.

and redundancy in social media data and extracted ADE reports with high precision.

4.7. Analysis of patient social media reports: a case on beta blocker

Beta blocker is the most discussed treatment in the test bed. It is a drug class consisting of multiple different treatments. There are 1822 discussions about beta blocker and its related medical events. Among them, 71% of them are adverse drug events, 20% are drug indications, and 9% are negated adverse drug events. Users often mentioned beta blocker with other treatments. Some of these treatments belong to the same drug class, others are co-medications. Fig. 5 shows the top 10 medications co-occurring with beta blocker in the heart disease discussion boards.

Among the top 10 co-occurred treatments, Atenolol, Metoprolol, Tenormin, Coreg and Inderal are beta blocker drugs. Ace inhibitor and calcium channel blocker are drug classes often used along with beta blocker to treat heart disease. Aspirin and Verapamil are not beta blocker treatments. Based on the analysis, we find that 50% of the adverse events related to beta blocker have other co-medications, presenting great potential for identifying drug interactions from these discussions.

FAERS and the forum place different emphases on ADEs. In Fig. 6, we compared the top 20 most discussed adverse Beta Blocker events from the forum with those from FAERS. Our system extracted 1297 patient-reported beta blocker ADEs; FAERS reported 3162. Among them, there are 5 common ADEs. FAERS focuses on severe ADEs, such as 'loss of consciousness' and 'death,' while forum reports concentrated on mild ADEs such as 'anxiety' and 'dizziness.' Forums seem to be more symptom derived describers while FAERS seems to be more diagnosis derived describers. Palpitations, fast heart rate and arrhythmia from forum discussions might be described as atrial fibrillation by the healthcare professionals.

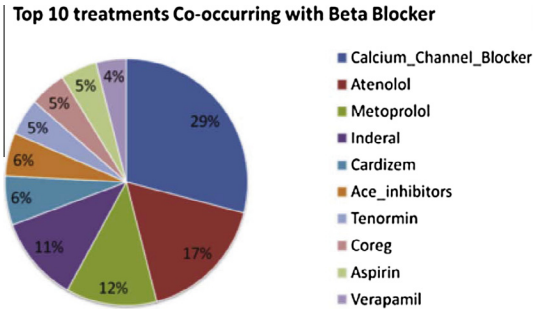

Fig. 5. Top 10 treatments co-occurring with beta blocker.

5. Conclusions and contributions

The advent of social media offers insights into healthcare unfiltered by traditional methods of healthcare data collection. Applying natural language processing techniques to extract patient reports of adverse drug events from social media has great potential to improve clinical and scientific knowledge of pharmacovigilance. In this study, we develop a research framework for pharmacovigilance in social media to identify patient reported adverse drug events. It 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.

A series of experiments were conducted on a test bed encompassing about one million postings. Our test bed includes diabetes discussions and heart disease discussions. The data came from 4 different sites and the performances are consistent. Our method can be generalized to different diseases and different types of discussion forums with minimal effort. The results reveal that each component of the framework significantly contributes to its overall effectiveness. Our proposed framework achieved an *f*-measure of about 85% in both the recognition of medical events and treatments. Our precision increased 40% in average and *f*-measure increased about 10% in adverse drug event extraction compared to methods in prior studies. The report source classification can effectively remove the noise in patient social media adverse event reports. Our framework significantly outperformed prior work in patient reported adverse drug event extraction.

The major contribution of our research is the design and evaluation of our research framework for pharmacovigilance research in health social media. It incorporates the state-of-the-art natural language processing techniques and effectively addresses the challenges in extracting patient adverse drug event reports from social media. This framework can be applied to analyze treatments of different diseases and extract patient intelligence on other medial related topics.

Conflict of interest

The authors declare that there are no conflicts of interest.

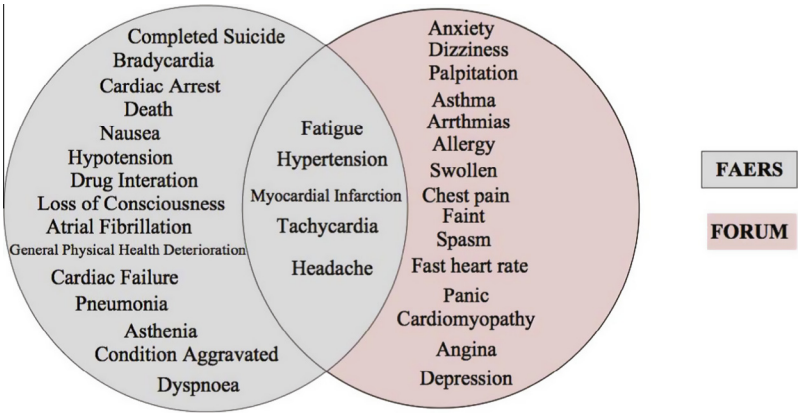

Fig. 6. A comparison of top 20 most reported adverse events for beta blocker from FAERS and forum.

Acknowledgements

This material is based upon work supported by a subcontract from Caduceus Intelligence Corporation with grant funds provided by the United States National Science Foundation (IIP-1417181).

We gratefully acknowledge the contribution of Dr. Randall Brown and Ms. Chanadda Chinthammit for their advices from clinical and pharmaceutical perspectives in this study. We thank Jing Liu for her analytical work on the MedHelp forum. We also appreciate the research assistance provided by fellow members of Healthcare Project team in University of Arizona's Artificial Intelligence Lab.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.10.011.

References

[1] A.R. Miller, C. Tucker, Active social media management: the case of health care, Inform. Syst. Res. 24 (1) (2013) 52–70.
[2] J.J. Mao, A. Chung, A. Benton, S. Hill, L. Ungar, C.E. Leonard, S. Hennessy, J.H. Holmes, Online discussion of drug side effects and discontinuation among breast cancer survivors, Pharmacoepidemiol. Drug Saf. 22 (3) (2013) 256–262.
[3] E. Basch, The missing voice of patients in drug-safety reporting, N. Engl. J. Med. 362 (10) (2010) 865–869.
[4] M. Hauben, A. Bate, Decision support methods for the detection of adverse events in post-marketing data, Drug Discov. Today 14 (7) (2009) 343–357.
[5] A. Bate, S. Evans, Quantitative signal detection using spontaneous ADR reporting, Pharmacoepidemiol. Drug Saf. 18 (6) (2009) 427–436.
[6] I.R. Edwards, M. Lindquist, Social media and networks in pharmacovigilance, Drug Saf. 34 (4) (2011) 267–271.
[7] R. Chunara, J.R. Andrews, J.S. Brownstein, Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak, Am. J. Trop. Med. Hyg. 86 (1) (2012) 39–45.
[8] R. Harpaz, W. DuMouchel, N.H. Shah, D. Madigan, P. Ryan, C. Friedman, Novel data-mining methodologies for adverse drug event discovery and analysis, Clin. Pharmacol. Ther. 91 (6) (2012) 1010–1021.
[9] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, 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, 2010, pp. 117–125.
[10] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C.E. Leonard, J.H. Holmes, Identifying potential adverse effects using the web: a new approach to medical hypothesis generation, J. Biomed. Inform. 44 (6) (2011) 989–996.
[11] A. Nikfarjam, G.H. Gonzalez, Pattern mining for extraction of mentions of adverse drug reactions from user comments, AMIA Annual Symposium Proceedings, vol. 2011, American Medical Informatics Association, 2011, p. 1019.
[12] A. Yates, N. Goharian, Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites, in: Advances in Information Retrieval, Springer, 2013, pp. 816–819.
[13] X. Liu, H. Chen, Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums, in: Smart Health, Springer, 2013, pp. 134–150.
[14] J. Bian, U. Topaloglu, F. Yu, Towards large-scale twitter mining for drug-related adverse events, in: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, ACM, 2012, pp. 25–32.
[15] A. Sarker, G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, J. Biomed. Inform. 53 (2015) 196–207.
[16] I. Segura-Bedmar, P. Martínez, R. Revert, J. Moreno-Schneider, Exploring spanish health social media for detecting drug effects, BMC Med. Inform. Decis. Making 15 (Suppl. 2) (2015) S6.
[17] B.W. Chee, R. Berlin, B. Schatz, Predicting adverse drug events from personal health messages, AMIA Annual Symposium Proceedings, 2011, American Medical Informatics Association, 2011, p. 217.
[18] H. Wu, H. Fang, S. Stanhope, et al., Exploiting online discussions to discover unrecognized drug side effects, Methods Inform. Med. 52 (2) (2013) 152–159.
[19] D.A. Lindberg, B.L. Humphreys, A.T. McCray, The unified medical language system, Methods Inform. Med. 32 (4) (1993) 281–291.
[20] J. Hadzi-Puric, J. Grmusa, Automatic drug adverse reaction discovery from parenting websites using disproportionality methods, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 792–797.
[21] M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, P. Bork, A side effect resource to capture phenotypic effects of drugs, Mol. Syst. Biol. 6 (1) (2010) 343.
[22] C.C. Yang, H. Yang, L. Jiang, M. Zhang, Social media mining for drug safety signal detection, in: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, ACM, 2012, pp. 33–40.
[23] H. Gurulingappa, L. Toldo, A.M. Rajput, J.A. Kors, A. Taweel, Y. Tayrouz, Automatic detection of adverse events to predict drug label changes using text and data mining techniques, Pharmacoepidemiol. Drug Saf. 22 (11) (2013) 1189–1194.
[24] Q.-C. Bui, S. Katrenko, P.M. Sloot, A hybrid approach to extract protein–protein interactions, Bioinformatics 27 (2) (2011) 259–265.
[25] K. Fundel, R. Küffner, R. Zimmer, Relex–relation extraction using dependency parse trees, Bioinformatics 23 (3) (2007) 365–371.
[26] Q. Zhu, R.R. Freimuth, J. Pathak, M.J. Durski, C.G. Chute, Disambiguation of PharmGKB drug–disease relations with NDF-RT and SPL, J. Biomed. Inform. 46 (4) (2013) 690–696.
[27] Z. Yang, H. Lin, Y. Li, Bioppisvmextractor: a protein–protein interaction extractor for biomedical literature using SVM and rich feature sets, J. Biomed. Inform. 43 (1) (2010) 88–96.
[28] I. Segura-Bedmar, P. Martinez, C. de Pablo-Sánchez, Using a shallow linguistic kernel for drug–drug interaction extraction, J. Biomed. Inform. 44 (5) (2011) 789–804.
[29] P. Thomas, M. Neves, I. Solt, D. Tikk, U. Leser, Relation extraction for drug–drug interactions using ensemble learning, Training 4 (2,402) (2011) 21–425.
[30] M. Miwa, R. Sætre, Y. Miyao, J. Tsujii, Protein–protein interaction extraction by leveraging multiple kernels and parsers, Int. J. Med. Inform. 78 (12) (2009) e39–e46.
[31] J. Li, Z. Zhang, X. Li, H. Chen, Kernel-based learning for biomedical relation extraction, J. Am. Soc. Inform. Sci. Technol. 59 (5) (2008) 756–769.
[32] D. Zelenko, C. Aone, A. Richardella, Kernel methods for relation extraction, J. Mach. Learn. Res. 3 (2003) 1083–1106.
[33] R.C. Bunescu, R.J. Mooney, A shortest path dependency kernel for relation extraction, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 724–731.
[34] F. Liu, L.D. Antieau, H. Yu, Toward automated consumer question answering: automatically separating consumer questions from professional questions in the healthcare domain, J. Biomed. Inform. 44 (6) (2011) 1032–1038.
[35] J. Huh, M. Yetisgen-Yildiz, W. Pratt, Text classification for assisting moderators in online health communities, J. Biomed. Inform. 46 (6) (2013) 998–1005.
[36] T. Fu, A. Abbasi, D. Zeng, H. Chen, Sentimental spidering: leveraging opinion information in focused crawlers, ACM Trans. Inform. Syst. (TOIS) 30 (4) (2012) 24.
[37] T. Joachims, Transductive inference for text classification using support vector machines, in: ICML, vol. 99, 1999, pp. 200–209.
[38] W. Chapman, Negex Version 2: A Simple Algorithm for Identifying Pertinent Negatives in Textual Medical Records, 2009.
[39] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (5) (2001) 301–310.
[40] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik, The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes, BMC Bioinform. 9 (Suppl. 11) (2008) S9.
[41] Ö. Uzuner, I. Goldstein, Y. Luo, I. Kohane, Identifying patient smoking status from medical discharge records, J. Am. Med. Inform. Assoc. 15 (1) (2008) 14–24.