Identifying Adverse Drug Events from Patient Social Media: A Case Study for Diabetes

Authors:

Xiao Liu, Department of Management Information Systems, University of Arizona

Hsinchun Chen, Department of Management Information Systems, University of Arizona

Abstract:

Patient social media sites have emerged as major platforms for discussions of treatments and drug side effects, making them a promising source for listening to patients' voices in adverse drug event reporting. However, extracting patient adverse drug event reports from social media continues to be a challenge in health informatics research. In light of the need for more robust extraction methods, we develop novel information extraction framework for identifying adverse drug events from patient social media. It consists of medical entity extraction for recognizing patient mentions of drugs and events, adverse drug event extraction using the shortest dependency path kernel-based statistical learning method and medical knowledge bases for semantic filtering, and report source classification to capture patient experience. A case study on a major diabetes patient social media platform is conducted to evaluate the performance. Our approach achieves an f-measure of 86% in the recognition of medical events and treatments, an f-measure of 69% for identifying adverse drug events, and an f-measure of 84% in patient report extraction. Our proposed methods significantly outperformed prior work in extracting patient reports of adverse drug events in health social media.

1. Introduction

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 drug effects (negative medical conditions occurring at the time a drug is used) or any other drug problem" [1]. Pharmacovigilance starts at the pre-approval stage, when information about adverse drug events (ADEs) is collected during phase I-III of clinical trials, and continues in the post-approval stage and throughout a drug's life on the market. Although clinical trials are used for evaluating safety issues, they are limited with respect to the number and characteristics of patients exposed, duration, and type of data collected. There are myriads of co-morbidities, over-the-counter and prescription drug interactions, and food interactions, which may take time to surface. The complete safety profile associated with a new drug cannot be fully established through clinical trials. Post-approval ADEs are a major health concern, accounting for more than 2 million injuries, hospitalizations, and deaths each year in the United States alone; associated costs are estimated at $75 billion annually [2]. Hence, timely drug safety surveillance after drugs' release to the market is an urgent goal of public health systems.

Recognizing the importance of drug safety surveillance, research into the identification, extraction, and detection of adverse drug events has steadily grown in the past decade. At the same time, social networks and patient forums on the Internet have emerged. Patient social media cover a large and diverse population and contain millions of unsolicited and uncensored discussions about medications. These discussions include information about drug indications (use of that drug for treating a particular medical condition) and adverse drug events (any medical condition or symptom occurring at the time a drug is used, whether or not it is identified as a cause of the injury). In particular, patient reports of adverse drug events in social media are more sensitive to underlying changes in patients' functional status than clinical and spontaneous reports. Thus, analyzing patient reports of adverse drug events in health social media may add value to current practice of pharmacovigilance by providing new perspectives for understanding drug effectiveness and side effects timely [3].

Given hundreds of health social network sites and forums available on the Internet, to identify patient reports of adverse drug events with manual approach is not feasible due to the scale of the problem and labor cost. In this paper, we seek to develop high-performance information extraction techniques for identifying patient reports of adverse drug events in health social media. A case study on a longitudinal diabetes patient social media platform is conducted to evaluate the performance. We believe that our approach is the first to combine statistical learning and semantic information for adverse drug event extraction. It captures adverse drug events based on patient experiences, providing an efficient way to listen to patients' voices for drug safety surveillance.

The remainder of this paper is organized as follows. Section 2 presents a review of current pharmacovigilance research with social media data. Section 3 describes our proposed research methods. Section 4 presents the evaluation methods and results. Section 5 concludes with a summary of our research contribution, practical implications, and future directions.

2. Related Work

There has been an increased interest in analyses of health social media content. We limit the scope of our review on prior pharmacovigilance studies to those that have used publically available social media data.

With respect to the test beds utilized, prior studies employed data sources from three different types of social media. Most studies utilized general health discussion forums, such as DailyStrength [4, 5], Yahoo health forums [6], and Medhelp [7]. General health social forums contain a variety of health-related topics ranging from herbal remedies to medications, thus filtering methods are necessary for extracting relevant information for subsequent analysis. Others developed research based on disease-focused discussion forums [3, 8]. Tweets (microblogs of 140 or fewer characters) have been employed in a recent study [9]. Among these data sets, disease-focused discussion forums are more suitable for adverse drug event detection as they contain more concentrated discussions about treatments for particular diseases [10].

A major objective of prior social media pharmacovigilance research is to extract adverse drug events [3-5, 9]. Chee et al. [6] utilized patient medication reviews to classify risky drugs and safe drugs for FDA scrutiny. Others explored the connections between adverse drug events and patient drug switching behaviors [8].

The most commonly used information extraction techniques are text classification, medical named entity recognition, and adverse drug event relation extraction. Classification methods such as support vector machines (SVM) and naïve Bayes have been applied in recent studies. Chee et al. [6] developed ensemble classifiers with SVM and naïve Bayes to classify risky drugs and safe drugs based upon online discussions. Bian et al. [9] used SVM to filter noise in tweets.

Medical named entity recognition in social media pharmacovigilance research aims to identify medically related entities (e.g., treatments and medical events). Most adopted lexicon-based entity recognition approaches due to the wide availability of medical lexicons and knowledge bases in the healthcare domain. The Unified Medical Language System (UMLS) has been adopted in prior studies [4, 5]. Spontaneous reporting systems (SRS), such as the FDA's Adverse Event Reporting System (FAERS) and MedEffect (adverse drug event reporting system in Canada), are often employed as a lexicon source [7, 9]. Since consumers' health vocabulary often differs from that of medical professionals [6], the Consumer Health Vocabulary, a lexicon linking UMLS standard medical terms to patients' colloquial language, has been adopted in many studies to interpret medical terms in online patient discussions [3, 7]. Nikfarjam et al. [5] developed a machine-learning-based association rule mining algorithm to generate patterns for recognizing adverse events.

After entity recognition, patient discussions of both drug and medical events can be extracted. Detecting whether a pair of drug and medical event is a report of an adverse drug event can be considered as a relation extraction task. The goal is to determine if there is a relation between the drug and events and the type of relation (e.g., drug indications or adverse drug events). Several prior studies have adopted co-occurrence analysis approaches to extract adverse drug event relations [3, 7, 8]. This approach assumes that if two entities are both mentioned within a certain range (e.g., within 20 tokens [5]), there is an underlying relationship between them.

In terms of results, several studies evaluated their performance using precision, recall, and f-measure metrics. For medical entity recognition, Leaman et al. [4] 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%. For relation extraction, all prior studies adopted co-occurrence analysis-based approaches [3, 7, 8]. None of these studies evaluated the performance since it is dependent on the data set. For text classification, Bian et al. [9] achieved 74% accuracy in identifying adverse events.

Based on our review of prior health social media pharmacovigilance research, we find lexicon-based approaches for medical entity extraction achieved better performance. Co-occurrence analysis-based adverse event extraction approach is widely adopted but has some clear drawbacks. There are multiple types of relations between medical events and drugs, including drug indications and adverse drug events. Sometimes patients negated connections between drugs and medical conditions in discussion. This approach, capturing little syntactic and semantic information in the sentences, may generate false adverse drug events when negations exist between medical events and drugs and confound adverse drug events with drug indications. The precision of co-occurrence analysis based approach is not sufficient to support further analysis on the extracted adverse drug events. It is critical to develop a more accurate adverse drug event extraction method in order to analyze patient reports of adverse drug events in social media.

Furthermore, although these studies extracted ADEs from patient forums, extracted ADEs may come from different report sources including patient experience, third-hand accounts, news and research. Most prior studies applied machine-learning-based classification techniques to filter out noise in health social media content. However, they rarely classified the adverse drug events based on the report sources to identify patient-reported ADEs, which have higher clinical value.

Our analysis of these studies motivated several critical directions that are incorporated in our case study, namely (1) the development and evaluation of a scalable and semantic-rich method for adverse drug event extraction and (2) a robust report source classification method to identify adverse drug events based on actual patient experience.

3. Methods

Our proposed research framework for identifying patient-reported adverse drug events is illustrated in Figure 1. Major components are explained in detail below.

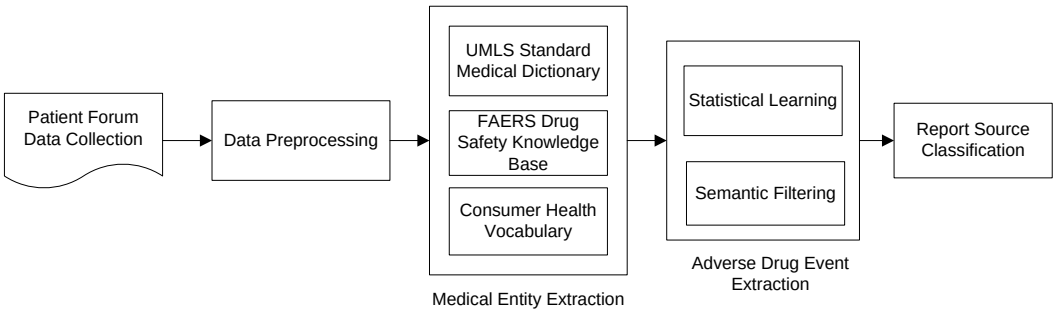

Fig 1. Research Framework for Identifying Patient-Reported Adverse Drug Events

3.1 Patient forum data collection

Diabetes affects 25.8 million people, or 8.3% of the American population. A large number of treatments exist to help control patients' glucose level and prevent organ damage from hyperglycemia. Many treatments have a number of adverse events that range from minor to serious, affecting patient safety to varying degrees. Patients' online discussions about their treatments can potentially give unique insights to drug safety surveillance and improve patient safety.

We developed our research test bed based on a major diabetes patient forum in the United States, the American Diabetes Association (ADA) online community (http://community.diabetes.org). An automated crawler was developed to download web pages and extract relevant 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. We collected 184,874 postings contributed by the ADA forum dating from Feb. 2009 (when the forum was established) to Dec. 2012 (our latest data collection).

3.2 Data preprocessing

Data preprocessing normalizes the raw data into a format that is ready for analysis. The preprocessing consists of two components: text cleaning and sentence boundary detection. In text cleaning, we developed specific regular expressions to remove URLs, duplicate punctuation, and personally identifiable information such as email address, social security number, etc. We then focus on sentence-level analysis in this study. We segment a post into sentences with a state-of-the-art open source natural language processing tool, OpenNLP (https://opennlp.apache.org). In total, there are 1,348,364 sentences in the test bed.

3.3 Medical entity extraction

We apply multiple types of lexicon sources to extract drug names and adverse events from the text, including UMLS, FAERS, and CHV, medical ontologies frequently used in prior studies [4-7, 9]. MetaMap (http://metamap.nlm.nih.gov), a highly configurable Java API from the National Library of Medicine, is used to map patient social media text to the UMLS [4]. We initialize the medical entity extraction with MetaMap to recognize terms matching standard medical lexicons in patient forums. Drug names and event names extracted by MetaMap are filtered by terms in the FDA's drug safety database, FAERS [7]. Terms that never appear in FAERS are not considered for further analysis. Then we extend the entity extraction to include the Consumer Health Vocabulary (CHV) [7]. For each term MetaMap identified, we query

the CHV to get its consumer-preferred equivalent and add it to our lexicon. The found consumer-preferred terms are used to search for additional entities in the patient forum. After the medical entity extraction, we identified 50,468 drug entities and 22,195 medical event entities. All sentences with both drug and event entities are extracted for further analysis. In total, we obtained unique 2972 sentences with at least one drug and one medical event.

3.4 Adverse drug event extraction

Patients' adverse drug event discussions in forums tend to be informal and colloquial, requiring medical knowledge and complex linguistic techniques to interpret. To address these issues, our approach incorporates statistical learning methods for relation detection and semantic information from medical and linguistic knowledge bases to identify adverse drug events from drug indications and negated ADEs.

3.4.1Statistical learning

An important task of adverse drug event extraction is to determine whether there is a relationship between a drug and medical event in a sentence. To detect related drug and medical events in patient forum posts, we developed a shortest dependency path kernel function and trained a Support Vector Machine (SVM) to learn patterns from posts with related drugs and events. Such kernel-based statistical learning methods have shown promise in identifying various relations in prior studies, such as protein interactions and drug interactions [11-13].

Feature generation.

We propose to generate syntactic and semantic features for relation instances based on the shortest dependency path from medical events to treatment entities. Dependency parsing captures both syntactic and semantic information between words in the sentences. It generates word-to-word links based on grammatical relations. In the dependency parse trees, the syntactic dependency is represented by the hierarchical structures of the trees. The semantic dependency is represented by the directions of the links. We utilized the Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml), which covers 53 different grammatical relations, for dependency parsing. A grammatical relation holds from a dependent to a governor (also known as a regent or a head). Figure 2 shows the dependency tree of a sample sentence.

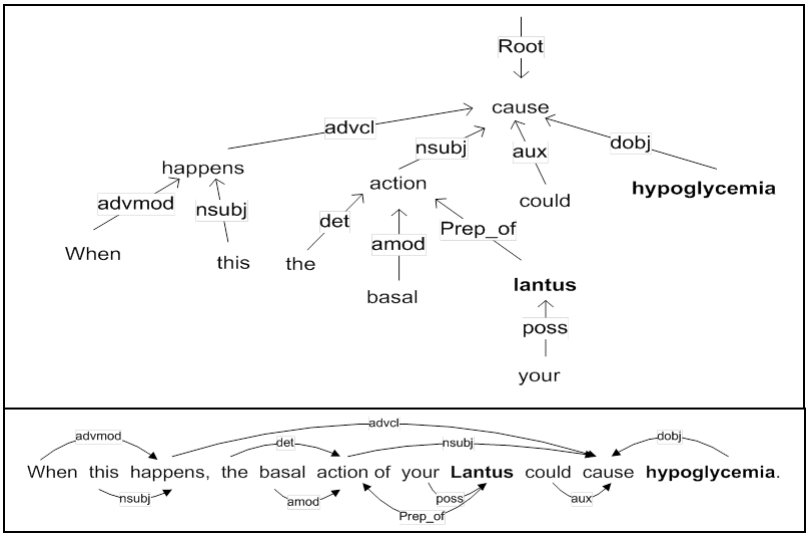

Fig 2. A sample sentence represented as a dependency tree

In this sentence, hypoglycemia is an adverse event entity and Lantus is a diabetes treatment. Grammatical relations between words are illustrated in the figure. For example, 'hypoglycemia' is the direct object of 'cause,' thus they have a grammatical relation 'dobj.' In this case, 'cause' is the governor and 'hypoglycemia' is the dependent. 'Action' is the noun subject of 'cause', thus they have a relation 'nsubj. '

Although the dependency tree presents the syntactic and semantic relationships between words in the sentences, a large proportion of the dependency tree is not relevant to the relationship between medication and medical event in the sentence. We utilized the shortest path between medical event entity and drug entity in the dependency tree (shortest dependency path) for feature generation.

Due to the large amount of data in the test bed, the representation of instances usually results in a large but sparse feature set, leading to decreased performance in training and testing. To reduce the data sparsity and increase the robustness of our method, we expand the shortest dependency path by categorizing words on the path into word classes with varying degrees of generality. Word classes include words, part-of-speech (POS) tags and generalized POS tags. POS tags are extracted with Stanford CoreNLP package (http://nlp.stanford.edu/software). We generalized POS tags according to Penn TreeBank guideline. Semantic types (Event and Treatments) are also used on the two ends of the shortest path.

The generated features for the relation instance hypoglycemia and Lantus can be defined as the Cartesian product of all the elements on the path, as illustrated in Figure 3. The original sentence thus can be represented in a sequence as X = [x_1, x_2, x_3, x_4, x_5, x_6, x_7], where x_1 = {Hypoglycemia, NN, Noun, Event}, x_2 = {->}, x_3={cause, VB, Verb}, x_4 = {<-}, x_5 = {action, NN, Noun}, x_6 = {<-}, x_7 = {Lantus, NN, Noun, Treatment}.

$$\begin{bmatrix} Hypoglycemia \\ NN \\ Noun \\ Event \end{bmatrix} \times [->] \times \begin{bmatrix} cause \\ VB \\ Verb \end{bmatrix} \times [<-] \times \begin{bmatrix} action \\ NN \\ Noun \end{bmatrix} \times [<-] \times \begin{bmatrix} Lantus \\ NN \\ Noun \\ Treatment \end{bmatrix}$$

Fig 3. Features generated from a dependency graph

Shortest dependency path kernel function.

Statistical learning methods rely on kernel functions to find a hyperplane that separates positive instances from negative. Given $x = x_1 x_2 x_3 ... x_m$ and $y = y_1 y_2 y_3 ... y_n$ are two relation instances, where x_i denotes the set of features corresponding to position i, the kernel function is computed as in the equation below:

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

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

For instance, 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}]. 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*1*1*1*2*1*3=18. Based on the result, we can see relation instances x and y have very high similarity scores. If instance x has a drug-event relation, instance y is very likely to contain a drug-event relation as well.

Classification.

We adopted Transductive Support Vector Machines (TSVM) [14] for classification in relation detection. Classification in relation detection is to distinguish relation instances with a relation from those without any relationship. Transductive Support Vector Machine (TSVM) is a semi-supervised machine-learning method that uses hyperplanes to find maximally distant separation between two classes of data based on kernel function. It can conduct learning with both

labeled and unlabeled data. SVM-light (http://svmlight.joachims.org), an open source package for TSVM is applied in this study because it supports customized kernel functions.

To conduct the statistical learning, we randomly selected 400 sentences with at least one drug entity and one medical event entity from each forum to serve as labeled data. We established content coding for labeling these sentences regarding whether the sentences contain related drug and medical event mentions. 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 it to identify instances with a drug-event relation. The procedures of statistical learning are summarized in Algorithm 1.

ALGORITHM 1. Statistical Learning Algorithm

Input: all relation instances with at least a pair of related drug and medical events, *R(drug, event).*
Output: where the instance has a pair of related drug and event.
Procedure:
1. **For** each relation instance *R(drug,event)* :
    Generate Dependency tree T of *R(drug,event)*
    Features = Shortest Dependency Path Extraction (T, R)
    Features = Syntactic and Semantic Classes Mapping (Features)
2. Separate relation instances into training set and test set
3. Train a SVM classifier C with shortest dependency kernel function based on the training set
4. Use the SVM classifier C to classify instances in the test set into two classes *R(drug, event)* = True and R(drug, event) = False.

3.4.2 Semantic filtering

Statistical learning methods can detect related drug and medical events but cannot precisely capture negation in sentences nor 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 low precision. To address these issues, we develop a semantic filtering algorithm, which utilizes the semantic knowledge from a drug safety databaseto remove drug indications and rules from the negation detection tool to filter out negated ADEs.

In the United States, the FDA strictly regulates indications for medications. Drug indications are well-documented in drug safety databases such as FAERS. We can obtain drug indication knowledge from existing knowledge bases such as FAERS to formulate templates and filter drug indications. For negation detection, we utilize the linguistic rule-based negation detection tool, NegEx [15]. NegEx is a natural language processing system for negation detection of medical events in medical documents. It is able to identify the negation phrases such as "never" and "no" and the scope of negation and then determine whether the medical events fall in the scope of negation. It has achieved 88% in precision and 85% in recall for identifying negated medical events. Given the adverse drug event in a sentence, we employ the NegEx to determine whether this event is negated or not. The detailed procedures for semantic filtering are presented in Algorithm 2.

ALGORITHM 2. Semantic Filtering Algorithm

Input: a relation instance i with a pair of related drug and medical events, *R(drug, event).*
Output: The relation type.
If drug exists **in** FAERS:
    Get indication list **for** drug;
    For indication **in** indication list:
        If event= indication:
            Return *R(drug, event)* = 'drug indication';
    For rule **in** NegEX:
        If relation instance i *matches* rule:
            Return *R(drug, event)* ='negated adverse drug events';
    Return *R(drug, event)* ='adverse drug events';

3.5 Report source classification

Reports of ADEs in social media may come from different report sources including patient experience, third-hand accounts, news and research. Among them, reports based on patient experiences have the most clinical value; others may introduce more noise and redundancy [2]. However, no previous patient social media research has differentiated patient reports of ADEs from third-hand accounts, news and research. To address this issue, report source classification is proposed to filter ADE reports not grounded in actual patients' experiences. We developed a feature-based classification model to distinguish patient reports from hearsay based on prior studies [10]. Bag-of-words (BOW) features and Transductive Support Vector Machines are utilized for report source classification.

To obtain training and evaluation data for classification, we randomly selected 400 sentences with at least one drug entity and one medical event entity from to create a gold standard evaluation dataset. 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. We represented instances in BOW features for report source classification. In total, we had 6,374 unique features. We applied the linear kernel in SVM-light for semi-supervised report source classification.

4. Evaluation and results

We use standard machine-learning and text analysis evaluation metrics, precision, recall, and f-measure, to evaluate the performances of our case study. These metrics have been widely used in information extraction and health social media studies.

$$\mathrm{Pr}ecision(i) = \frac{\#\text{ of correctly identified instances for class i}}{\mathrm{Total}\,\#\text{ of instances identified as class i}}$$

$$\mathrm{Re}call(i) = \frac{\#\text{ of correctly identified instances for class i}}{\mathrm{Total}\,\#\text{ of instances in class i}}$$

$$F-measure(i) = \frac{2 * precision(i) * recall(i)}{precision(i) + recall(i)}$$

To evaluate the performance of medical entity extraction, we randomly selected 200 sentences from the test data and established definitions and content coding for labeling entities and medical event entities. 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. These sentences became the gold standard for entity recognition evaluation. We then compared the results from our automatic tagger against the gold standard. To evaluate our approach for adverse drug event extraction with both statistical learning and semantic filtering (SL+SF), we established content coding for labeling adverse drug events based on information in existing knowledge bases and advice from clinical experts. 400 sentences with at least one drug entity and one medical event entity were randomly selected and annotated to serve as the gold standard for evaluation. There are 762 relation instances in the gold standard dataset, including 302 instances with no related drug and event, 276 adverse drug event relations, 15 negated adverse drug event relations and 169 drug indication relations. To justify the selection of kernel function in statistical learning, we compared the results from shortest dependency path (SDP) kernel with bag-of-words (BOW) kernel. We compared extraction results from our approach against the gold standard. We report the performance of combining statistical learning and semantic filtering to extract adverse drug events and compare it results without semantic filtering. To demonstrate the efficacy of our proposed method, we conducted co-occurrence (CO) 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 within 20 tokens of an event term, then this was treated as a co-occurrence [3].

We conducted 5-fold cross validation to obtain the evaluation results for adverse drug event extraction and report source classification. 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. Table 1 below summarizes our evaluation results.

Component	Category	Precision	Recall	F-measure
Medical Named Entity Extraction	Drug	93.90%	91.70%	92.50%
	Medical Event	87.30%	80.30%	83.50%
Adverse Drug Event Extraction	SL with BOW	27.34%	77.36%	40.36%
	SL with SDP	61.50%	59.81%	60.64%
	SL+SF approach	81.70%	59.81%	69.06%
	CO approach	36.22%	100%	53.18%
Report Source Classification	With RSC	83.50%	84.10%	83.80%
	Without RSC	62.50%	100%	76.92%

Table 1: Evaluation results of our research framework

For significance testing, we created two contingency tables for adverse drug event extraction and report source classifications based on the results of 5 fold cross validations over 762 instances. Fisher's Exact Test was adopted to compute the p values for null hypotheses as shown in Table 2. Both p values are below 0.01. The associations between methods and outcomes are significant for both adverse drug event extraction and report source classification.

Adverse Drug Event Extraction	Accurate ADE	Inaccurate ADE
SL+SF	163	38
CO	276	486
P value <0.01		

Report Source Classification	Accurate patient report	Inaccurate patient report
With RSC	399	78
Without RSC	476	286
P value <0.01		

Table 2: Contingency tables for Fisher's Exact Test

Our approach achieved 93.9% in precision, 91.7% in recall, and 92.5% in f-measure for drug entity extraction. Regarding medical event entity extraction, our precision is 87.3%, recall is 80.3%, and f-measure is 83.5%. The effectiveness of our medical entity extraction procedure helps significantly with our subsequent analyses. Based on the evaluation results, we observe that our approach significantly increases the precision and f-measure for adverse drug event extraction. The shortest dependency path kernel outperformed the bag-of-words kernel. Our method achieved 82% in precision, 60% in recall, and 69% in f-measure. The co-occurrence baseline method achieves 36% in precision, 100% in recall, and 53% in f-measure. The f-measure of our approach is about 16% higher. Without report source classification (RSC), the performance of extraction is heavily affected by noise in the discussion. The precision is 62.5%, recall is 100%, and f-measure is 76.9% without RSC. After report source classification, the precision increased to 83.5% and the overall performance (f-measure) increased to 83.8%. Applying the proposed techniques in our test bed, we obtained 1069 adverse drug event relations and among them 652 are patient reports. It took each rater 10-15 hour effort to create a gold standard dataset with 400 sentences. It is time consuming and costly to review health social media manually for patient reports of adverse drug events given the scale of the problem. Compared to manual approach, our proposed approach minimized the manual effort and managed to improve the efficiency of patient social media adverse drug event report extraction. Compared to the baseline methods, our approach significantly improved the accuracy and overall quality of the social media adverse drug event reports, which provides more reliable evidence for risky drug identification.

5. Conclusion

In light of the need for more robust adverse drug event extraction methods, we develop a research framework for enhanced patient-reported ADE extraction. It consists of medical entity extraction for recognizing patient discussions of

drugs and events and adverse drug event extraction using the shortest dependency path kernel-based statistical learning method, medical knowledge bases for semantic filtering, and report source classification to capture patient experience.

A case study was conducted on a test bed created from a major diabetes forum in the United States. The results reveal that each component significantly contributes to its overall effectiveness. Our proposed approach achieved an overall f-measure of 86% in both the recognition of medical events and treatments. Our f-measure increased 16% in adverse drug event extraction compared to methods in prior studies. 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 a novel approach for identifying adverse drug events in patient social media. It integrates the state-of-the-art techniques from different domains and effectively addresses the challenges in extracting patient adverse drug event reports from social media. This approach may contribute to drug safety regulation by providing more reliable evidence for earlier risky drug identification.

ACKNOWLEDGMENTS

This work was supported in part by DTRA, #HDTRA1-09-1-0058. We gratefully acknowledge the contribution of Randall Brown, MD, a physician at University Medical Center University of Arizona, to this study. We also appreciate the research assistance provided by fellow members of the Smart Health Project team in the University of Arizona's Artificial Intelligence Lab.

Reference

[1] Hauben, M., & Bate, A. (2009). Decision support methods for the detection of adverse events in post-marketing data. Drug discovery today, 14(7), 343-357.

[2] Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics, 91(6), 1010-1021.

[3] Benton A., Ungar L., Hill S., Hennessy S., Mao J., Chung A., & Holmes J. H. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. Journal of biomedical informatics, 44(6), pp. 989-996.

[4] Leaman R., Wojtulewicz L, Sullivan R, Skariah A., Yang J, Gonzalez G. (2010) 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, ACL, pp.117-125.

[5] Nikfarjam A., Gonzalez G.H. (2011). Pattern mining for extraction of mentions of Adverse Drug Reaction from user comments. In: Proceeding of 2011 AMIA annual Symposium, pp. 1019-1026.

[6] Chee B. W., Berlin R., & Schatz B. (2011). Predicting adverse drug events from personal health messages. In: AMIA Annual Symposium Proceedings Vol. 2011, pp. 217-226.

[7] Yang C. C., Yang H., Jiang L., & Zhang M.: Social media mining for drug safety signal detection. In: Proceedings of the 2012 international workshop on Smart health and wellbeing ACM pp. 33-40 (2012).

[8] Mao, J. J., Chung, A., Benton, A., Hill, S., Ungar, L., Leonard, C. E., Holmes, J. H. (2013). Online discussion of drug side effects and discontinuation among breast cancer survivors, (January), 256–262.

[9] Bian, J., Topaloglu, U., & Yu, F. (2012, October). Towards large-scale twitter mining for drug-related adverse events. In Proceedings of the 2012 international workshop on Smart health and wellbeing (pp. 25-32). ACM.

[10] Liu, X., & Chen, H. (2013). AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In Smart Health (pp. 134-150). Springer Berlin Heidelberg.

[11] Qian, L., & Zhou, G. (2012). Tree kernel-based protein–protein interaction extraction from biomedical literature. Journal of Biomedical Informatics, 45(3), 535-543.

[12] Bunescu R.C., Mooney R.J. (2005). A Shortest Path Dependency Kernel for Relation Extraction. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 724-731.

[13] Miwa, M., Sætre, R., Miyao, Y., & Tsujii, J. I. (2009). Protein–protein interaction extraction by leveraging multiple kernels and parsers. International journal of medical informatics, 78(12), e39-e46.

[14] Joachims T. (1999). Transductive inference for text classification using support vector machines. In: machine learning- international workshop pp. 200-209.

[15] Chapman, W. W. (2009). NegEx version 2: a simple algorithm for identifying pertinent negatives in textual medical records.

[16] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums, ACM Transactions on Information Systems, 26(3) no. 12

The authors' bios

Xiao Liu is a doctoral student in the Department of Management Information Systems and a research associate in the Artificial Intelligence Lab at University of Arizona. Her research interests include health informatics, machine learning and social media analytics. Contact here at xiaoliu@email.arizona.edu.

Hsinchun Chen is the University of Arizona Regents' Professor and Thomas R. Brown Chair in Management and Technology in the Department of Management Information Systems and the funding director of the Artificial Intelligence Lab. His research interest include Web computing, search engines, digital libraries, intelligence analysis, biomedical informatics, data/text/Web mining, and knowledge management. Chen has a PhD in information systems from New York University. He is a Fellow of IEEE and AAAS. Contact him at hchen@eller.arizona.edu.

Contact Info

Xiao Liu

Phone: 520-288-2889

Address: McClelland Hall, Room 430.1130 E. Helen Street, Tucson, AZ 85719

Email: xiaoliu@email.arizona.edu

Hsinchun Chen

Phone: 520-621-2748

Address: McClelland Hall 430X, 1130 E. Helen Street, Tucson, AZ 85721

Email: hchen@eller.arizona.edu