Identifying Adverse Drug Events from Patient Social Media

A Case Study for Diabetes

Xiao Liu, *University of Arizona*

Hsinchun Chen, *University of Arizona and Tsinghua University*

Social media sites have emerged as major platforms for discussion of treatments and drug side effects, making them a promising source for adverse drug event reporting, but extracting such reports remains challenging.

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 preapproval stage, when information about adverse drug events (ADEs) is collected during phases 1 through 3 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 evaluation safety issues, they're limited with respect to the number and characteristics of patients exposed, duration, and type of data collected. Myriad co-morbidities, over-the-counter and prescription drug interactions, and food interactions can take time to surface, thus the complete safety profile associated with a new drug can't 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 US alone, with associated costs estimated at $75 billion annually.[2] Timely safety surveillance after a drug's release on the market is therefore an urgent goal of public health systems.

Recognizing the importance of drug safety surveillance, research into the identification, extraction, and detection of ADEs has steadily grown in the past decade. At the same time, social networks and patient forums on the Internet have emerged and increased in popularity as evidenced by site traffic. 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 ADEs (any medical condition or symptom occurring at the time a drug is used, whether or not it's identified as a cause of the injury). In particular, patient reports of ADEs through social media are more sensitive to underlying changes in patients' functional status than clinical and spontaneous reports. Thus, analyzing these reports of ADEs in health social media could

1541-1672/15/$31.00 © 2015 IEEE
Published by the IEEE Computer Society

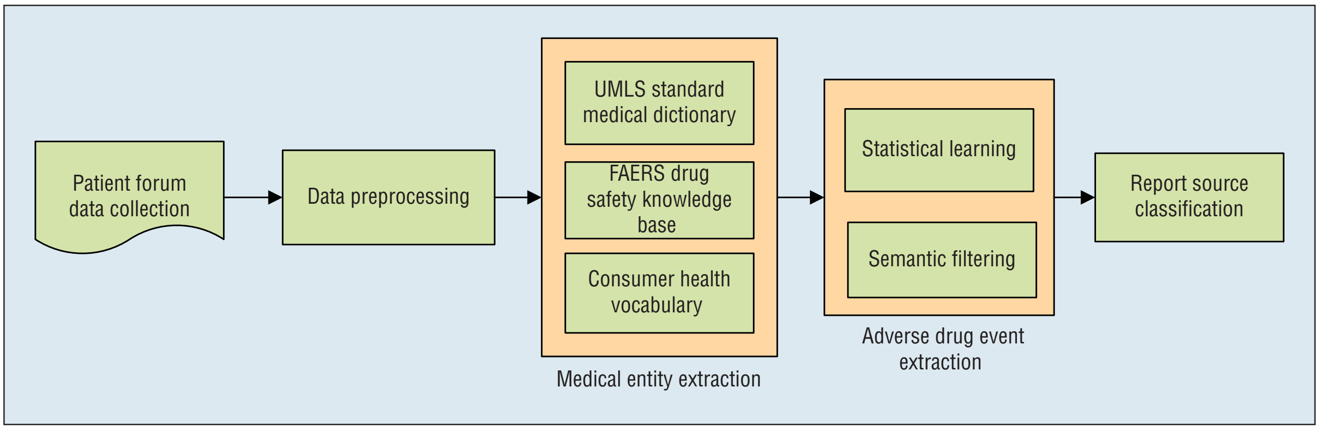
Patient forum data collection
Data preprocessing
UMLS standard medical dictionary
FAERS drug safety knowledge base
Consumer health vocabulary
Medical entity extraction
Statistical learning
Semantic filtering
Adverse drug event extraction
Report source classification

Figure 1. Research framework for identifying patient-reported adverse drug events. UMLS stands for Unified Medical Language System (UMLS) and FAERS stands for the FDA's Adverse Event Reporting System. Our research framework consists of five major components: patient forum data collection, data preprocessing, medical entity extraction, adverse drug event extraction, and report source classification.

add value to the current practice of pharmacovigilance by providing new perspectives for understanding drug effectiveness as well as side effects.[3]

Given the hundreds of health social network sites and forums available on the Internet, manually identifying patient reports of ADEs isn't feasible, so we describe high-performance automatic information extraction process. A case study on a longitudinal diabetes patient social media platform evaluates our approach's performance. We believe that ours is the first approach to combine statistical learning and semantic information for ADE extraction (see the sidebar for related work). Specifically, it captures ADEs based on patient experiences, providing an efficient way to listen to patients' voices during drug safety surveillance.

Methods

Figure 1 illustrates our proposed research framework for identifying patient-reported ADEs; in the following, we explain the major components in detail.

Patient Forum Data Collection

Diabetes affects 25.8 million people, or 8.3 percent 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 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 testbed based on a major diabetes patient forum in the US, the American Diabetes Association (ADA) online community (http://community.diabetes.org). An automated crawler was developed to download webpages 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 February 2009 (when the forum was established) to December 2012 (our latest data collection).

Data Preprocessing

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

Medical Entity Extraction

We apply multiple types of lexicon sources to extract drug names and adverse events from the text, including the Unified Medical Language System (UMLS), the US Food and Drug Administration's (FDA's) Adverse Event Reporting System (FAERS), and the Consumer Health Vocabulary (CHV), medical ontologies frequently used in prior studies.[4-9] 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 the standard medical lexicons in patient forums. Drug 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 aren't considered for further analysis.

Next, we extend the entity extraction to include the CHV.[7] For each term that MetaMap identified, we query the CHV to get its consumer-preferred equivalent and add it to our lexicon. The consumer-preferred terms are then 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 and extracted those sentences with both drug and event entities for further analysis. In total, we obtained 2,972 unique sentences with at least one drug and one medical event.

ADE Extraction

Patients' ADE 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 ADEs from drug indications and negated ADEs.

Statistical learning. An important task of ADE extraction is to determine whether there's 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.[10-12]

We propose generating 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, generating word-to-word links based on grammatical relations. In the dependency graph, syntactic dependency is represented by the hierarchical structures of the trees, and semantic dependency is represented by the links' directions. We used 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

Related Work Using Social Media for Pharmacovigilance

Although there has been an increased interest in analyses of health social media content, we limit our scope to those prior pharmacovigilance studies that have used publicly available social media data.

With respect to testbeds, prior studies employed data sources from three different types of social media. Most studies accessed *general health discussion forums*, such as DailyStrength,[1,2] Yahoo health forums,[3] and Medhelp.[4] *General health social forums* contain a variety of health-related topics ranging from herbal remedies to medications, thus filtering methods are necessary to extract relevant information for subsequent analysis. Others developed research based on *disease-focused discussion forums*.[5,6] Tweets (microblogs of 140 or fewer characters) drove a recent study.[7] Among these datasets, disease-focused discussion forums are more suitable for adverse drug event (ADE) detection because they contain more concentrated discussions about treatments for particular diseases.[8]

A major objective of prior social media pharmacovigilance research is to extract ADEs.[1,2,5,7] Brant Chee and his colleagues used patient medication reviews to classify risky versus safe drugs for US Food and Drug Administration (FDA) scrutiny.[3] Others explored the connections between ADEs and patient drug-switching behaviors.[6]

The most commonly used information extraction techniques are text classification, medical named entity recognition, and ADE relation extraction. Classification methods such as support vector machines (SVMs) and naïve Bayes have been applied in recent studies. Chee and colleagues[3] developed ensemble classifiers with SVM and naïve Bayes to classify risky and safe drugs based on online discussions, whereas Jiang Bian and his colleagues used SVM to filter noise in tweets.[7]

Medical named entity recognition in social media pharmacovigilance research aims to identify medically related entities (both treatments and medical events). Most studies adopted lexicon-based entity recognition approaches because of the wide availability of medical lexicons and knowledge bases in the healthcare domain. Prior studies used the Unified Medical Language System (UMLS), for example.[1,2] Spontaneous reporting systems (SRSs), such as the FDA's Adverse Event Reporting System (FAERS) and MedEffect (the ADE reporting system in Canada), often are used as a lexicon source.[4,7] Because consumers' health vocabulary often differs from that of medical professionals,[3] 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.[4,5] Azadeh Nikfarjam and his colleagues developed a machine-learning-based association rule mining algorithm to generate patterns for recognizing adverse events.[2]

Using medical named entity recognition, researchers can extract patient discussions of both drug and medical events. The system then treats this data as a relation extraction task, detecting whether a pair of drug and medical events is a report of an ADE. The goal is to determine if there's a relation between the drug and events and the type of relation (for example, drug indications or ADEs). Several prior studies have adopted co-occurrence analysis approaches to extract ADE relations.[4-6] This approach assumes that if two entities are both mentioned within a certain range (say, within 20 tokens[2]), there's an underlying relationship between them.

In terms of results, several studies evaluated performance using precision, recall, and f-measure metrics. For medical entity recognition, Robert Leaman and his colleagues achieved the best performance values on extracting adverse events from forums with a precision of 78.3 percent, recall of 69.9 percent, and f-measure of 73.9 percent.[1] For relation extraction, all prior studies adopted co-occurrence analysis-based approaches.[4-6] None of these studies evaluated the performance because it's dependent on the dataset. For text classification, Bian and colleagues achieved 74 percent accuracy in identifying adverse events.[7]

Based on our review of prior health social media pharmacovigilance research, we find that lexicon-based approaches for medical entity extraction achieved better performance. The co-occurrence analysis-based adverse event extraction approach is widely adopted, but it has some clear drawbacks. For example, there are multiple types of relations between medical events and drugs, including drug indications and ADEs. Patients sometimes negated the connections between drugs and medical conditions in their discussion. This approach, capturing little syntactic and semantic information in the sentences, could generate false ADEs when negations exist between medical events and drugs or confound ADEs with drug indications. The precision of a co-occurrence analysis-based approach therefore isn't sufficient to support further analysis on extracted ADEs. Instead, we need a more accurate ADE extraction method to analyze patient reports via social media.

Furthermore, although these studies extracted ADEs from patient forums, they could 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 ADEs based on 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, the development and evaluation of a scalable and semantic-rich method for ADE extraction and a robust report source classification method to identify ADEs based on actual patient experience.

References

1. R. Leaman et al., "Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks," *Proc. 2010 Workshop Biomedical Natural Language Processing*, 2010, pp. 117–125.
2. A. Nikfarjam and G.H. Gonzalez, "Pattern Mining for Extraction of Mentions of Adverse Drug Reaction from User Comments," *Proc. 2011 AMIA Ann. Symp.*, 2011, pp. 1019–1026.
3. B.W. Chee, R. Berlin, and B. Schatz, "Predicting Adverse Drug Events from Personal Health Messages," *Proc. 2011 AMIA Ann. Symp.*, 2011, pp. 217–226.
4. C.C. Yang et al., "Social Media Mining for Drug Safety Signal Detection," *Proc. 2012 Int'l Workshop Smart Health and Wellbeing*, 2012, pp. 33–40.
5. A. Benton et al., "Identifying Potential Adverse Effects Using the Web: A New Approach to Medical Hypothesis Generation," *J. Biomedical Informatics*, vol. 44, no. 6, 2011, pp. 989–996.
6. J.J. Mao et al., "Online Discussion of Drug Side Effects and Discontinuation among Breast Cancer Survivors," *Pharmacoepidemiology and Drug Safety*, Jan. 2013, pp. 256–262.
7. J. Bian, U. Topaloglu, and F. Yu, "Towards Large-Scale Twitter Mining for Drug-Related Adverse Events," *Proc. 2012 Int'l Workshop Smart Health and Wellbeing*, 2012, pp. 25–32.
8. X. Liu and H. Chen, "AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums," *Smart Health,* Springer, 2013, pp. 134–150.

a governor (also known as a regent or head). Figure 2 shows the dependency graph of a sample sentence.

In this sentence, *hypoglycemia* is an adverse event entity and *Lantus* is a diabetes treatment. The figure shows the grammatical relations between words, such as *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 isn't relevant to the relationship between medication and medical event in the sentence. We used 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 testbed, the representation of instances usually results in a large but sparse feature set, leading to decreased performance in training and testing. To reduce data sparsity and increase robustness in 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 the Stanford CoreNLP package (http://nlp.stanford.edu/software); we generalized them according to the Penn TreeBank guideline. Semantic types (event and treatments) are also used on the two ends of the shortest path.

We can define the generated features for the relation instance *hypoglycemia* and *Lantus* as the Cartesian product of all the elements on the path, as Figure 3 shows. We can thus represent the original sentence 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 = {<-}, and x_7 = {Lantus, NN, Noun, Treatment}.

Statistical learning methods rely on kernel functions to find a hyperplane that separates positive instances from negative. Given that $x = x_1\ x_2\ x_3 \ldots xm$ and $y = y_1\ y_2\ y_3 \ldots yn$ are two relation instances, where x_i denotes

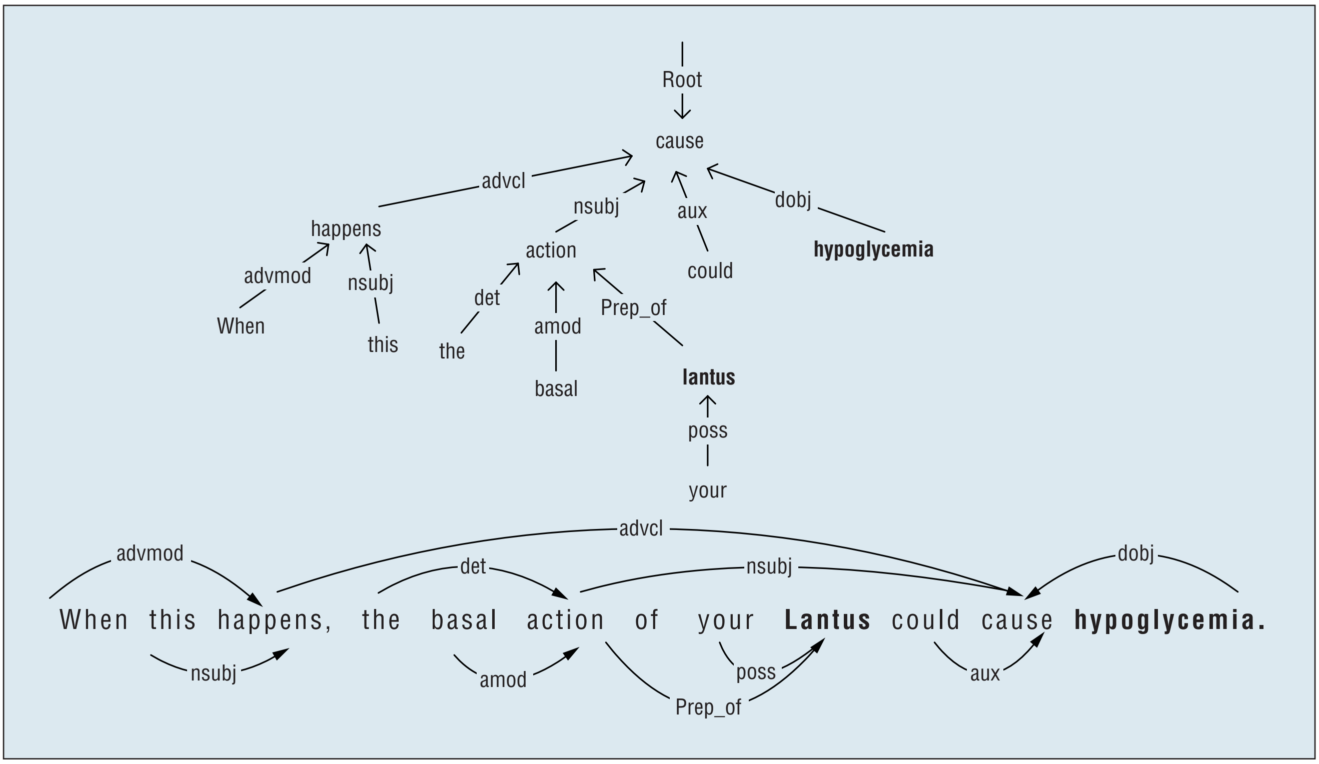

Figure 2. A sample sentence represented as a dependency graph. Each node on the graph is a word, with the links between nodes representing semantic dependency between words.

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

Figure 3. Features generated from a dependency graph. These features tend to be word, entity types, POS tags, and generalized POS tags on the shortest path between two entities on the dependency graph.

the set of features corresponding to position i, the kernel function is computed as follows:

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

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

So, we can represent instance x = {When this happens, the basal action of your Lantus could cause hypoglycemia.} as x = [{Hypoglycemia, NN, Noun, Event}, {->}, {cause, VB, Verb}, {<-}, {action, NN, Noun}, {<-}, {Lantus, NN, Noun, Treatment}]. We can represent instance y = {But, now I've read a few posts in this thread that indicate depression as a possible side effect from Lantus.} as y = [{depression, NN, Noun, Event}, {->}, {indicate, VBP, Verb}, {<-}, {effect, NN, Noun}, {<-}, {Lantus, NNP, Noun, Treatment}]. The system can then compute $K(x, y)$ as the product of the number of common features xi and yi in position i, thus, $K(x, y) = 3*1*1*1*2*1*3 = 18$. 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.

We adopted transductive SVMs (TSVMs)[13] for classification in relation detection, which helps distinguish relation instances with a relation from those without any relationship. TSVM is a semisupervised 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 and 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 and trained the TSVM classifier on the shortest dependency path kernel before applying it to identify instances with a drug-event relation. Figure 4 summarize the procedures for statistical learning.

```
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 an 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.
```

Figure 4. Statistical learning algorithm. It takes in relation instances, trains an SVM classifier on syntactic and semantic features, and predicts whether a pair of drug and event entities has a relation.

Semantic filtering. Statistical learning methods can detect related drug and medical events but can't precisely capture negation in sentences or differentiate drug indication relations from ADEs. 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 developed a semantic filtering algorithm that utilizes the semantic knowledge from a drug safety database to remove drug indications and rules from the negation detection tool to filter out negated ADEs.

```
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 matchesrule:
        Return R(drug, event) = 'negated adverse drug events';
  Return R(drug, event) = 'adverse drug events';
```

Figure 5. Semantic filtering algorithm. It takes a pair of related drug events and classifies the relation based on semantic rules generated from FAERS and NegEX.

In the US, 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 use the linguistic rule-based negation detection tool, NegEx,[14] a natural language processing system for negation detection of medical events in medical documents. It can identify 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 percent in precision and 85 percent in recall for identifying negated medical events. Given the ADE in a sentence, we can use NegEx to determine whether this event is negated. Figure 5 gives the detailed procedures for semantic filtering.

Report Source Classification

Reports of ADEs in social media can come from different 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 can 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.[15] Bag-of-words (BOW) features and TSVMs assist in 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 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. Two research associates were trained to label the selected sentences from each forum based on these rules. In total, we had 6,374 unique

Table 1. Evaluation results of our research framework.*

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
ADE 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	100.00	53.18
Report source classification	With RSC	83.50	84.10	83.80
	Without RSC	62.50	100.00	76.92

* ADE = adverse drug event, BOW = bag of words, CO = co-occurrence, RSC = report source classification, SDP = shortest dependency path, SF = semantic filtering, SL = statistical learning.

features. We applied the linear kernel in SVM-light for semisupervised report source classification.

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:

$$\text{Precision}(i) = \frac{\begin{array}{c}\text{\# of correctly identified}\\ \text{instances for class } i\end{array}}{\begin{array}{c}\text{Total \# of instances}\\ \text{identified as class } i\end{array}}$$

$$\text{Recall}(i) = \frac{\begin{array}{c}\text{\# of correctly identified}\\ \text{instances for class } i\end{array}}{\begin{array}{c}\text{Total \# of instances}\\ \text{in class } i\end{array}}$$

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

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. We then compared the results from our automatic tagger against this gold standard.

To evaluate our approach for ADE extraction with both statistical learning and semantic filtering (SL+SF), we established content coding for labeling ADEs based on information in existing knowledge bases and advice from clinical experts. Four hundred 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: we had 762 relation instances in the gold standard dataset, including 302 instances with no related drug and event, 276 ADE relations, 15 negated ADE relations, and 169 drug indication relations. To justify the selection of kernel function in statistical learning, we compared the results from the shortest dependency path (SDP) kernel with the BOW kernel. We then compared extraction results from our approach against the gold standard. To demonstrate the efficacy of our proposed method, we conducted co-occurrence (CO) analysis-based ADE extraction as a baseline for comparison. We adopted this approach from a prior study, in which if a drug occurred within 20 tokens of an event term, it was treated as a co-occurrence.[3]

We conducted five-fold cross-validation to obtain the evaluation results for ADE extraction and report source classification—each time, we used 80 percent of labeled data and all the unlabeled sentences in our testbed as a training set and 20 percent of labeled data as a test set. Table 1 summarizes our evaluation results.

For significance testing, we created two contingency tables for ADE extraction and report source classifications based on the results of the five-fold cross-validations over 762 instances—specifically, we adopted Fisher's Exact Test to compute the p values for null hypotheses (see Table 2). Both p values are below 0.01. The associations between methods and outcomes are significant for both ADE extraction and report source classification.

Our approach achieved 93.9 percent in precision, 91.7 percent in recall, and 92.5 percent in f-measure for drug entity extraction. Regarding medical event entity extraction, our precision was 87.3 percent, recall 80.3 percent, and f-measure 83.5 percent. Based on the evaluation results, we observe that our approach significantly increases the precision and f-measure for ADE extraction. The SDP kernel outperformed the BOW kernel, and our method achieves 82 percent in precision, 60 percent in recall, and 69 percent in f-measure. In contrast, the CO baseline method achieves 36 percent in precision, 100 percent in recall, and 53 percent in f-measure.

Without report source classification (RSC), the extraction performance is heavily affected by noise in the discussion—specifically, the precision is 62.5 percent, recall is 100 percent, and f-measure is 76.9 percent without RSC. After RSC, the precision increased to 83.5 percent, and overall performance (f-measure) increased to 83.8 percent. Applying the proposed techniques in our testbed, we obtained 1,069 ADE relations and among them, 652 are patient reports. It took each rater 10 to 15 hours of effort to create a gold standard dataset with 400 sentences.

Compared to a fully manual approach, our proposed method minimized manual effort and managed to improve efficiency. Compared to baseline methods, our approach significantly improved the accuracy and overall quality of the social media ADE reports, which provides more reliable evidence for risky drug identification.

In the future, we're going to examine the uniqueness and novelty of patient reported ADE in social media. We plan to develop a meta-learning method to aggregate ADEs from multiple sources and predict drugs with high risks of adverse events earlier. ■

Table 2. Contingency tables for Fisher's Exact Test.

ADE 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		

THE AUTHORS

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

Hsinchun Chen is the University of Arizona Regents' Professor and Thomas R. Brown Chair in Management and Technology in its Department of Management Information Systems as well as the funding director of the Artificial Intelligence Lab. He's also the National 1000-Elite Program Chair Professor at the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, China. His research interests 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's a Fellow of IEEE and AAAS. Contact him at hchen@eller.arizona.edu.

Acknowledgments

This work was supported in part by the US Defense Threat Reduction Agency (DTRA) grant HDTRA1-09-1-0058 and China NSFC project # 71110107027. We gratefully acknowledge the contribution of Randall Brown, a physician at the 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.

References

1. M. Hauben and A. Bate, "Decision Support Methods for the Detection of Adverse Events in Post-Marketing Data," *Drug Discovery Today*, vol. 14, no. 7, 2009, pp. 343–357.
2. R. Harpaz et al., "Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis," *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, 2012, pp. 1010–1021.
3. A. Benton et al., "Identifying Potential Adverse Effects Using the Web: A New Approach to Medical Hypothesis Generation," *J. Biomedical Informatics*, vol. 44, no. 6, 2011, pp. 989–996.
4. R. Leaman et al., "Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks," *Proc. 2010 Workshop Biomedical Natural Language Processing*, 2010, pp. 117–125.
5. A. Nikfarjam and G.H. Gonzalez, "Pattern Mining for Extraction of Mentions of Adverse Drug Reaction from User Comments," *Proc. 2011 AMIA Ann. Symp.*, 2011, pp. 1019–1026.
6. B.W. Chee, R. Berlin, and B. Schatz, "Predicting Adverse Drug Events from Personal Health Messages," *Proc. 2011 AMIA Ann. Symp.*, 2011, pp. 217–226.
7. C.C. Yang et al., "Social Media Mining for Drug Safety Signal Detection," *Proc. 2012 Int'l Workshop Smart Health and Wellbeing*, 2012, pp. 33–40.
8. J.J. Mao et al., "Online Discussion of Drug Side Effects and Discontinuation among Breast Cancer Survivors," *Pharmacoepidemiology and Drug Safety*, Jan. 2013, pp. 256–262.
9. J. Bian, U. Topaloglu, and F. Yu, "Towards Large-Scale Twitter Mining for Drug-Related Adverse Events," *Proc. 2012 Int'l Workshop Smart Health and Wellbeing*, 2012, pp. 25–32.
10. L. Qian and G. Zhou, "Tree Kernel-Based Protein–Protein Interaction Extraction from Biomedical Literature," *J. Biomedical Informatics*, vol. 45, no. 3, 2012, pp. 535–543.
11. R.C. Bunescu and R.J. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 724–731.
12. M. Miwa et al., "Protein–Protein Interaction Extraction by Leveraging Multiple Kernels and Parsers," *Int'l J. Medical Informatics*, vol. 78, no. 12, 2009, pp. e39–e46.
13. T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," *Proc. Int'l Workshop Machine Learning*, 1999, pp. 200–209.
14. W. Chapman et al., "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries," *J. Biomedical Informatics*, vol. 34, no. 5, 2001, pp. 301–310.
15. X. Liu and H. Chen, "AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums," *Smart Health*, Springer, 2013, pp. 134–150.