Postprint: Research on the Construction of an ADR Signal Detection Model Integrating Statistical Learning and Semantic Filtering

Authors: Wei Wei, Zheng Du

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] The emergence of social media provides new avenues for collecting medical and health data. Applying natural language processing techniques to extract patient-reported ADR (Adverse Drug Reaction) signals from social media holds great potential for improving clinical and scientific knowledge in adverse drug reaction monitoring. However, extracting patient-reported ADR signals from social media still faces significant challenges. To this end, we develop a research model that utilizes advanced natural language processing techniques to extract ADR signals from health-themed social media. [Methods/Process] The model first employs a multi-dictionary source matching-based approach to identify medical entities from noisy social media; then extracts adverse drug reactions using a statistical learning method based on the shortest dependency path kernel function; utilizes semantic knowledge from drug safety databases to filter out drug therapeutic and indication information as well as negated adverse drug reactions; and finally classifies reporting sources to eliminate rumors and other noise information. [Results/Conclusion] The effectiveness of the model is validated by collecting data from diabetes forums, and the results demonstrate that each component of the model contributes to the improvement of its overall performance.

Full Text

A Study on Adverse Drug Reaction Signal Extraction Framework Based on Integrated Statistical Learning and Semantic Filter

Wei Wei[1], Zheng Du[2]

[1]Big Data Institute, Zhongnan University of Economics and Law, Wuhan 430074
[2]School of Information Management, Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] The emergence of social media provides a new approach for collecting healthcare data. Applying natural language processing techniques to extract patient-reported Adverse Drug Reaction (ADR) signals from social media holds great potential for improving the clinical and scientific knowledge of ADR monitoring. However, extracting ADR signals from patient reports in social media remains a major challenge. This paper proposes a research framework for extracting ADR signals from health-related social media using advanced natural language processing techniques. [**Method/Process**] The framework first employs a multi-dictionary source matching approach to identify medical entities from noisy social media data. It then uses statistical learning methods based on shortest dependency path kernels to extract adverse drug events. Semantic knowledge from drug safety databases is utilized to filter out drug treatment and indication information, as well as negated adverse drug events. Finally, report sources are classified to remove noise such as rumors. [**Result/Conclusion**] Data collected from diabetes forums were used to validate the model's effectiveness, with results demonstrating that each component of the model contributes to its overall performance improvement.

Keywords: medical entity recognition; adverse drug event extraction; health social media; statistical learning; semantic filtering

1. Introduction

In recent years, with the rapid development of the Internet and Web 2.0-based social media, the way people access health information has gradually changed. From passively communicating medical information face-to-face with healthcare professionals in the past, to now actively searching for and sharing health information through health-themed social media, individuals hope to participate in the daily management of their own health. An increasing number of patients are willing to share their diagnosis, treatment, medication, and side effect information, as well as their emotional experiences in battling diseases, on the Internet—especially in online health communities. This makes such platforms unique and powerful sources for obtaining health, medication, and treatment information. Patients' self-reports on social media often include medical issues and adverse reactions that clinicians might miss or overlook. The accumulated comment information in online health communities represents first-hand data from medication users, containing rich potential ADR information. Social media is considered a new channel for collecting drug side effects and treatment effects, as it can enhance the acquisition of subjective elements regarding drug safety and treatment management, providing important insights for clinical practice.

Given the clinical and scientific value of patient-reported content on social media, researchers have begun exploring methods to identify and extract ADR signals from these platforms. However, extracting high-quality patient-reported content from this noisy environment remains challenging. Drug adverse events

are medically sound descriptions of medical conditions caused by medications. Negated drug adverse events represent denials of causal relationships between drugs and adverse events. Drug indications are fundamental medical knowledge about medications used to treat diseases. In patients' online narratives, treatment information and medical events frequently appear mixed together, and discussions may include drug treatment information, indication information, and negated drug adverse events. ADR signals from social media may originate from patients' real experiences or from research, news, rumors, or duplicated information, resulting in report sources containing substantial noise and redundant data. Table 1 illustrates these phenomena through posts from diabetes online communities.

As shown in Table 1, online community users employ their preferred healthcare language in discussions, which differs from professional medical terminology. For example, in post 63828, "stroke" is a user-preferred expression, while the term in FAERS (FDA Adverse Event Reporting System) would be "cerebrovascular" (cerebrovascular disease). In post 34188, "bruising" is expressed as "contusion" in FAERS. Additionally, patients' discussions may contain different types of drug-event relationships. In post 63828, the author mentions "stroke" and "Lipitor"—a lipid-lowering agent that reduces stroke risk—representing a drug-indication relationship rather than a drug-adverse reaction relationship. In post 9043, the patient reports chest pain while taking "Actos" (a hypoglycemic drug), representing a drug adverse event. Forum information may also come from different report sources: post 63828 concerns diabetes research, posts 9043, 25139, and 34188 are personal medication experience comments, while post 12200 is hearsay from others.

2. Related Research

Medical entity recognition aims to identify medical entity objects such as treatments and drugs. Thanks to rich medical dictionaries and knowledge bases in the healthcare domain, many previous studies have adopted dictionary-based entity recognition methods. UMLS (Unified Medical Language System developed by the U.S. National Library of Medicine) is frequently employed in research. Spontaneous reporting systems are also commonly used as data sources for extracting treatments and adverse events from text. FAERS medical terminology is often used to map drug and adverse event entities from health social media. MedEffect (Canada's adverse drug event reporting system) has also been used to extract adverse events from social media. However, user-generated healthcare language on health social media differs from professional medical terminology. Due to individual knowledge and preference differences, expressions in drug reviews may be diverse and varied.

Biomedical relation extraction techniques have been used to identify relationships such as gene-disease relationships and protein interactions from free text. Adverse drug event extraction employs relation extraction techniques to determine whether relationships exist between drugs and events and to identify

relationship types (e.g., drug-indication or drug-adverse reaction relationships). Adverse drug reaction relation extraction methods can be divided into three categories: co-occurrence analysis-based methods, rule-based methods, and statistical learning-based methods. These methods differ in how they utilize lexical, syntactic, and semantic information from text. Rule-based syntactic and semantic information extraction methods demonstrate good performance. Statistical learning-based relation extraction methods can automatically learn relation patterns from annotated corpora, making them more suitable for large-scale corpora. Supervised statistical learning dominates entity relation extraction. Among these, kernel-based entity relation extraction is a representative approach. Kernel-based relation extraction methods have shown advantages in identifying various biomedical relationships such as protein interactions and gene-disease relationships. P. Thomas et al. employed composite kernels integrating graph kernels and shortest dependency path kernels to extract drug-drug interaction relationships from medical literature. These methods utilize syntactic and semantic information to more concisely and accurately capture relationships between entities, achieving better results than feature-based relation extraction methods. This approach uses kernel functions to integrate multiple aspects of syntactic and semantic information, with final entity relation distances composed of multiple kernel functions from different information sources, thereby improving accuracy.

Methods for identifying medical entities from preprocessed data include rule-based, dictionary-based, and statistical learning approaches. In practice, one or several methods are typically selected based on specific task requirements to achieve better recognition performance. Literature surveys by S. Abed et al. show that ADR dictionaries and knowledge bases have been widely used data resources for ADR signal extraction from social media. These biomedical data resources contain ADR lists collected from drug labels to clinical trials, caregivers, and even user posts on social media. Most previous studies have used precision, recall, and F-value metrics to evaluate performance. To demonstrate the value of adverse drug reactions reported from social media, researchers have conducted multiple analyses on extraction results. A. Benton et al. compared adverse events extracted from social media with recorded adverse drug events, finding that extracted events could achieve 35.1% precision, 77% recall, and 52.8% F-value. B. Chee et al. found that patient drug comments could be used to identify risky drugs on the market, with most identified risky drugs appearing on the FDA's drug safety watch list. C. Yang et al. considered health social media an important data source with broad application prospects for ADR signal detection.

Through literature review, we found that medical entity extraction based on medical dictionaries and ontologies can achieve satisfactory results. However, few studies have applied advanced statistical learning methods for relation extraction to mine adverse drug reactions from health social media. Co-occurrence analysis-based adverse drug reaction extraction methods have obvious limitations: they cannot effectively capture syntactic or semantic information, poten-

tially resulting in incorrect adverse reaction relationships when negation exists in sentences; extracted drug adverse events may be confused with drug indications; and when multiple drug-adverse event entities appear in the same sentence, this method cannot accurately capture relationships between drugs and adverse reactions. Furthermore, health social media contains many repeated reports of news, research, and stories from third-party accounts, generating redundancy and noise that reduces the accuracy of ADR signal identification from social media—an issue rarely addressed in previous studies. As an increasingly prominent open platform where users can freely express their questions and demands through web communities, the value of health social media remains far from fully exploited.

3. ADR Signal Extraction Framework Integrating Statistical Learning and Semantic Filtering

The research question addressed in this paper can be described as: developing an integrated and scalable research framework for mining patient-reported ADR signals from health-themed social media; identifying genuine patient-reported content from noisy user discussions; and demonstrating that statistical learning methods require enhanced health-related semantic filtering to improve adverse drug event extraction results compared to baseline methods.

The proposed framework includes four components: medical entity recognition based on multi-dictionary sources, adverse drug event relation extraction based on shortest dependency path kernels, semantic filtering using medical knowledge base information, and report source classification to filter out rumors and other noise information, as shown in Figure 1 [Figure 1: see original paper].

3.1 Data Collection and Preprocessing Data preprocessing typically involves cleaning and standardizing data to prepare raw data for subsequent analysis. The preprocessing stage includes two steps: text cleaning and sentence segmentation. Based on regular expressions, URLs, repeated punctuation marks, and personally identifiable information such as email addresses, personal accounts, and phone numbers are removed from the text. This eliminates irrelevant information while retaining useful information to ensure the speed and quality of iterative processes. The proposed method focuses on information extraction and processing at the sentence level. Therefore, the natural language processing tool OpenNLP is used to segment each post into independent sentences. OpenNLP provides state-of-the-art machine learning-based sentence boundary detection algorithms for this purpose.

3.2 Medical Entity Recognition Based on Multi-Dictionary Sources Extracting medical entities from noisy user-generated content is a challenging task. R. Leaman et al.'s dictionary-based research has proven to be the best-performing medical entity recognition system. Dictionary-based methods rely on existing dictionaries, typically using string matching or similarity calculations

to identify drug and adverse event entities from free text. The performance of this recognition method depends on the comprehensiveness of the underlying reference dictionary and the quality of the similarity algorithm. This study utilizes UMLS, FAERS, and CHV as multi-dictionary sources to extract drug names and adverse drug events from social media text.

3.2.1 Standard Medical Entity Extraction Based on UMLS MetaMap is a Java API linked to the U.S. National Library of Medicine, used to identify UMLS medical concepts from health social media. Currently, UMLS has 135 semantic types, which are further abstracted into 15 semantic groups such as "Chemicals and Drugs," "Disorders," and "Genes & Molecular Sequences." By configuring MetaMap, drug name entities belonging to the "Chemicals and Drugs" semantic group and adverse drug event entities belonging to the "Disorders" semantic group are identified. MetaMap first identifies medical concepts in patient comments that match the standard medical dictionary UMLS.

3.2.2 Medical Vocabulary Filtering Based on FAERS MetaMap mapping results for the "Chemicals and Drugs" and "Disorders" semantic groups may contain some false positive information. For example, food and formula ingredients in forum discussions are often identified as belonging to the "Chemicals and Drugs" semantic group, while common verbs such as "find" and "have" may be extracted as belonging to the "Disorders" semantic group. To avoid these issues, FDA's FAERS is used to filter drug and adverse event names extracted by MetaMap, removing medical entity names that do not appear in FAERS for further analysis.

3.2.3 Consumer Health Vocabulary Expansion for Online Health Terms Drug adverse events discussed by patients in online health communities differ from those in biomedical literature or clinical notes, as these discussions typically contain more informal and colloquial expressions that require medical knowledge and complex language technology for parsing. Patient discussions in online health communities often include user-preferred medical vocabulary and descriptive text. To more comprehensively and accurately understand patient discussions on social media, this study integrates Consumer Health Vocabulary (CHV) as a dictionary source to expand the richness of online user expressions. CHV contains 47,505 UMLS standard medical terms and corresponding 127,081 user-preferred terms. For each medical entity retained from the previous step, CHV is queried to obtain its corresponding user-preferred terms, which could not previously be recognized by MetaMap. These user-preferred terms are then used to retrieve the patient comment dataset to expand medical entity extraction. Integrating user-preferred vocabulary mentioned in online health communities further expands the extracted medical entity set.

3.3 Adverse Drug Event Relation Extraction Based on Shortest Dependency Path Kernel Through the review of biomedical relation extraction research, combined with statistical learning methods for relation detection and semantic filtering based on medical knowledge, this study proposes a statistical learning method based on shortest dependency path kernels to extract adverse drug events. This method leverages existing medical knowledge and statistical learning techniques to significantly enhance the accuracy of extracted adverse events.

3.3.1 Feature Generation Drug adverse events discussed in online health communities typically contain substantial colloquial expressions. Due to data sparsity, statistical learning based on lexical and distance features yields unsatisfactory results. However, patient narratives about drug adverse events still follow certain syntactic and semantic rules. Therefore, this study proposes extracting syntactic and semantic features from sentence dependency parse trees to represent instances. Dependency parsing generates word-to-word links based on syntactic relationships, representing grammatical and semantic information between words in a sentence. In dependency parse trees, syntactic dependencies are displayed in the tree's hierarchical structure, while semantic dependencies are shown in the direction of links. The Stanford Parser is used for dependency parsing to extract grammatical relationships from dependents to governors. The Stanford Parser employs context-free grammar and lexicalized dependency parsing to generate dependency relationships between components in the dependency tree. Figure 3 [Figure 3: see original paper] shows a dependency tree for a sentence where "nausea" is an adverse event entity and "Byetta" is a diabetes treatment drug. The figure illustrates grammatical relationships between words—for example, "nausea" is the direct object of "gotten," so they have the grammatical relationship "dobj." In this case, "gotten" is the governor and "nausea" is the dependent.

Most of the dependency tree is irrelevant to drug and medical condition relationships in the sentence. Previous research has shown that the contribution of dependency trees to establishing relationships between two entities is almost entirely concentrated on the shortest path between them. To utilize the shortest path between drug entities and medical event entities in dependency relationships (shortest dependency path), an algorithm is proposed to extract the shortest path between two entities from the dependency tree. The algorithm searches for the shortest path from the medical event to the drug treatment in the dependency tree for each relation instance, capturing not only words but also the direction of dependencies on the path. The shortest dependency path extraction process is described in Pseudo-code 1:

Pseudo-code 1:
Input: A relation instance i, a pair of related drug and adverse event R(drug, event) = True, dependency graph T
Output: Path, the shortest dependency path from event to drug

Procedure: ShortestDependencyPathExtraction(i, drug, event, T)
1. if drug.event.dependents() then
2. Path ← {event.←, drug}
3. else
4. Path ← {event}, End ← {drug}, Head ← {event}, Tail ← {drug}
5. while Head ≠ Tail.governor do
6. if drug Head.dependents() then
7. Head ← Head.governor
8. Path ← Path + {→.Head, ←, drug}
9. else
10. Head ← Head.governor
11. Path ← Path + {→.Head}
12. if event Tail.governor.dependents() then
13. Tail ← Tail.governor
14. End ← {event, →, Tail, ←} + End
15. else
16. Tail ← Tail.governor
17. End ← {Tail, ←} + End
18. Path ← Path + {←} + End
19. return path

3.3.2 Grammatical and Semantic Class Mapping To increase the robustness of the extraction method, words on the path are part-of-speech (POS) tagged to extend the shortest dependency path. After shallow syntactic parsing of sentences, the Stanford CoreNLP software package is used to extract and annotate POS information, following the Stanford Penn TreeBank guidelines. Semantic types (event and drug treatment) are labeled at both ends of the shortest path. Table 2 lists the POS tags involved in the dataset.

Table 2. Part-of-Speech (POS) Tags

POS Tags	Description
CC	Conjunction (连词)
CD	Cardinal number (基数)
DT, PDT	Determiner (限定词)
IN	Preposition (介词)
JJ, JJR, JJS	Adjective (形容词)
NN, NNS, NNP, NNPS	Noun (名词)
PRP, PRPS	Pronoun (代词)
RB, RBR, RBS	Adverb (副词)
RP	Particle (小品词)
UH	Interjection (感叹词)
VB, VBD, VBG, VBN, VBZ, VBP	Verb (动词)
WDT, WP, WPS, WRB	Wh-words (特殊疑问词)
EX, FW, LS, MD, SYM	Others

Matching patterns and kernel functions are defined based on nodes in the syntactic parse tree. Matching patterns reflect whether the POS tags and features of two nodes match. Kernel functions are used for matching and similarity calculation of relation instances to determine relationships between entities. The feature representation of a relation instance can be defined as the Cartesian product of all elements on the path. For the example sentence in Figure 3, the feature representation is:

X = {x_1, x_2, x_3, x_4, x_5}, where
x_1 = {Nausea, NN, Noun, Event},
x_2 = {→},
x_3 = {gotten, VBD, Verb},
x_4 = {←},
x_5 = {Byetta, NN, Noun, Treatment}

Nausea_{Event} × [→] × gotten_{Verb} × [←] × Byetta_{Treatment}

3.3.3 Shortest Dependency Path Kernel Kernel-based methods can utilize various data organization forms to represent entity relationships. When calculating distances between relationships, kernel functions are used instead of inner products of feature vectors. Kernel functions implicitly compute the dot product of objects' feature vectors in high-dimensional feature spaces—in many cases, the product of the number of common features at their locations can be calculated without enumerating all features. Statistical learning methods rely on kernel functions to find a hyperplane that separates positive instances from negative instances. For the shortest dependency path kernel, if $x = x_1x_2x_3x_4…x$ and $y = y_1y_2y_3y_4…y$ are two relation instances, where x represents the feature set corresponding to position i, the kernel function is defined as:

$K(x, y) = \Sigma_{1}\ C(x , y)$ (1)

C(x , y) = |x y | is the number of common features between x and y . Whether two relation instances have the same relationship type can be determined through kernel function calculation. The shortest dependency path kernel pseudo-code is described in Pseudo-code 2:

Pseudo-code 2:
Input: Relation instances $x = x_1x_2…x$ and $y = y_1y_2…y$
Output: K(x, y), similarity score between x and y
Procedure: ShortestDependencyPathKernel(x, y)
1. If m ≠ n then
2. K(x, y) ← 0
3. else
4. while i ≤ m do
5. K(x, y) ← K(x, y) × |x y |
6. return K(x, y)

For example, consider relation instance x = "When this happens, the basal

action of your Lantus could cause hypoglycemia." with feature representation x = [{Hypoglycemia, NN, Noun, Event}, {→}, {cause, VB, Verb}, {←}, {action, NN, Noun}, {←}, {Lantus, NNP, Noun, Treatment}]. Another relation instance y = "But now I've read a few posts in this thread that indicated 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}]. The kernel function K(x, y) calculates the product of the number of common features at position i. In this example, K(x, y) = $3 \times 1 \times 1 \times 1 \times 2 \times 1 \times 3 = 18$. Based on this result, relation instances x and y have a very high similarity score. If relation instance x contains a certain drug-event relationship, relation instance y is also likely to contain that drug-event relationship.

3.3.4 Classification Classification in entity relation detection aims to separate instances with certain relationship features from those without such features. Transductive Support Vector Machine (TSVM) is used for relation detection classification. SVM-light is an open-source software package that supports TSVM, has been adopted in previous research with good results, and importantly, supports user-defined kernel functions. The TSVM classifier is trained using the shortest dependency path kernel, which is then applied to identify drug-event relation instances. The detailed process is described in Pseudo-code 3:

Pseudo-code 3:
Input: All relation instances I, each containing at least one drug-event pair
Output: Whether a drug-event pair is related, R(drug, event) = True or False
Procedure: StatisticalLearningAlgorithm(drug, event)
1. For each drug-event pair, R(drug, event) do
2. Generate dependency graph T containing R(drug, event) instance i
3. Path ← ShortestDependencyPathExtraction(i, drug, event, T)
4. Feature ← GrammaticalAndSemanticClassMapping(Path)
5. Divide relation instances into training and test sets
6. Train an SVM classifier C on the training set using the shortest dependency path kernel
7. Use classifier C on the test set to classify relation instances into two categories:
R(drug, event) = True
R(drug, event) = False

3.4 Semantic Filtering The shortest dependency path kernel can detect related drug-adverse reaction relationships, but it cannot precisely capture negation relationships in sentences or distinguish drug indications from adverse drug reactions. Previous studies have neglected the importance of filtering drug indications and negated adverse drug reactions, resulting in low accuracy of extracted adverse drug events. To address this issue, this study employs a semantic filtering method based on semantic knowledge from drug safety databases to filter out drug indication information and uses negation detection tools to

filter out negated adverse drug event information (as shown in the lower part of Figure 2 [Figure 2: see original paper]).

3.4.1 Drug Indication Labeling Based on FAERS When patients share medication experiences or comment on drugs in communities, they inevitably mention reasons for medication use or drug indications. For example, a comment about Metoprolol states: "I use this primarily for my hypertension." In this statement, "hypertension" is the reason for medication use, not an adverse reaction of Metoprolol. Since drug indications are standardized and detailed records exist in drug safety databases (e.g., FAERS), drug indication knowledge is obtained from FAERS to form templates. MetaMap is used to identify relevant biomedical entities from FAERS indication descriptions, filtering out drug indication information mixed with adverse drug events.

3.4.2 Negated Adverse Drug Event Filtering Based on NegEx For negated adverse drug event detection, the language rule-based negation detection tool NegEx is employed. NegEx is a natural language processing system previously used for detecting negated medical events in discharge summaries. It has been used to annotate biomedical texts and identify medical events from discharge records. The semantic filtering process is described in Pseudo-code 4:

Pseudo-code 4:
Input: Relation instance i with a drug-event pair R(drug, event) = True
Output: T(drug, event), relationship type between drug and event
Procedure: SemanticFilteringAlgorithm(drug, event)
1. if drug FAERS.drug() then
2. indications ← FAERS.indication(drug)
3. if event indications then
4. return T(drug, event) = Drug Indication
5. for rule NegEx do
6. if instance i matches rule then
7. return T(drug, event) = Negated Adverse Drug Event
8. else
9. return T(drug, event) = Adverse Drug Event

3.5 Report Source Classification To reduce noise and redundancy in patient-reported adverse drug reactions, report source classification is used to filter out drug adverse event reports not based on actual patient experiences. Online health communities contain news, stories, rumors, and other duplicated or reposted messages related to adverse drug reactions from third-party accounts. These messages are not actual adverse drug reaction experiences from patients. Previous health social media research has not addressed this issue. Literature review reveals that text classification techniques can effectively identify healthcare professional posts in Yahoo! Answers and recognize drug users from tweets. These tasks are similar to identifying actual patient experiences of adverse drug events from online health communities. Statistical

learning-based classification techniques can help filter out this noise data from social media.

To classify report sources of adverse drug events in social media, a feature-based classification model is used to distinguish patient reports from rumors. Bag-of-Words (BOW) features and Transductive Support Vector Machine (TSVM) are employed to classify news, research, rumors, and duplicated report sources, distinguishing actual patient-reported adverse drug events from hearsay. TSVM builds models using both labeled and unlabeled data, performing transductive inference on unlabeled data with a small set of labeled data.

4. Experiments and Results Analysis

4.1 Dataset and Preprocessing Chronic diseases such as diabetes and heart disease often rely on patient self-management. The emergence of many online health forums provides platforms for chronic disease patients to communicate anonymously, where they can consult questions, acquire knowledge, and share their emotional experiences in disease treatment. The experimental dataset in this study is sourced from the well-known U.S. diabetes community Diabetes-Forums, with the interface shown in Figure 4 [Figure 4: see original paper]. DiabetesForums is a large diabetes support online community with over 50,000 registered users. Most site users are diabetes patients, with some being diabetes caregivers. The community gathers the latest news and discussions on diabetes symptoms, treatments, monitoring, diet, and research.

The Octopus web scraper was used to crawl 67,444 posts from DiabetesForums between January 1, 2009, and December 31, 2015. Since the proposed method focuses on information extraction and processing at the sentence level, the natural language processing tool OpenNLP was used for sentence segmentation, yielding 42,355 sentences.

4.2 Evaluation Metrics Standard statistical learning and text analysis evaluation metrics—precision, recall, and F-value—are used to assess framework performance. These metrics have been widely applied in information extraction and health social media research. Co-occurrence analysis methods are commonly used for adverse drug event relation extraction due to their simplicity, making them a benchmark for comparison with other improved methods. Following literature [11], if a drug and an adverse event co-occur 20 times or more in a post, they are considered co-occurring.

4.3 Results Analysis

4.3.1 Medical Entity Recognition First, MetaMap is used to identify UMLS medical concepts from the health social media dataset. FAERS is then used for filtering, removing medical entity names not appearing in FAERS. For each retained medical entity, CHV is queried to obtain corresponding user-preferred terms that could not previously be recognized by MetaMap. These

user-preferred terms are used to retrieve the patient comment dataset to expand medical entity extraction. Table 3 shows the medical entity recognition results using the proposed method.

Table 3. Medical Entity Recognition Performance

Method	Drug Entity	Medical Event
	Precision	Recall
Our Method	0.92	0.88

The results show that the F-value for drug entity extraction using the proposed method reaches 90%, while the F-value for medical event extraction exceeds 80%. The good performance is primarily attributed to the combination of consumer health vocabulary, knowledge-based filtering, and the FAERS drug safety database. Additionally, since drugs and medical events discussed in diabetes forums are all diabetes-related, terminology consistency is higher compared to other health communities with diverse backgrounds and topics, contributing to higher performance.

Analysis of experimental results reveals that errors in drug entity recognition mainly stem from spelling errors and abbreviations of drug names. Medical event entity recognition shows lower performance than drug entity recognition because errors primarily originate from patients' vague descriptions of medical events. For example, patient descriptions of "hypo-symptoms" and "a low" both refer to hypoglycemia, but these vague descriptions cannot be identified during actual extraction. To further improve performance, more advanced machine learning-based named entity taggers need to be applied.

4.3.2 Adverse Drug Event Relation Extraction For drug-adverse reaction relation detection, 400 sentences were randomly selected from the dataset. The proposed method focuses on identifying drug and medical event relationships within the same sentence; cross-sentence drug-medical event relationships within the same post are not addressed in this study.

Based on information from existing knowledge bases and clinical expert recommendations, content coding annotation was performed on these sentences according to the proposed method. Each drug-medical event pair in a sentence is treated as a relation instance. Two researchers annotated these relation instances, with disagreements resolved by a third party. Statistical information of annotated relation instances is shown in Table 4 .

Table 4. Statistics of Drug-Event Relations in Annotated Dataset

Relation Type	Count
Adverse Drug Event	237

Relation Type	Count
Drug Indication	89
Negated Adverse Drug Event	41
No Relation	33
Total	400

To demonstrate the effectiveness of the proposed method, it is compared with co-occurrence analysis-based adverse drug event extraction methods. The performance results of three different methods for adverse drug event extraction are shown in Table 5 .

Table 5. Performance Results of Three Different Methods for Adverse Drug Event Extraction

Method	Precision (%)	Recall (%)	F1 (%)
Co-occurrence (CO)	52.3	100.0	68.7
Statistical Learning (SL)	73.8	60.2	66.4
SL + Semantic Filtering (SL+SF)	83.1	60.2	69.8

The comparison results show that the proposed method can significantly improve the precision and F-measure of adverse drug event extraction. Statistical learning helps improve precision while causing a decrease in recall. Semantic filtering further improves precision without affecting recall. The precision of the proposed method is about 31% higher than co-occurrence analysis, and the F-measure is about 10% higher. The precision of co-occurrence analysis mainly depends on dataset quality. Since users in health communities discuss not only drug treatment effects but also diverse topics such as diagnosis, symptoms, drug indications, and reasons for medication use, and their discussions sometimes involve numerous drug names, the precision of co-occurrence analysis is relatively low. However, for pharmacovigilance research, accurately capturing ADR signals is more meaningful than obtaining large numbers of false reports. The proposed method can improve the accuracy of ADR signal extraction from social media, enhancing the quality of adverse drug event reports from health social media.

We also note that using the shortest dependency path kernel-based statistical learning method results in decreased recall (from 100% to about 60%). This is caused by errors in relation detection for long-sentence relation instances, which appear less frequently in annotated data, leading to low learning rates and recall. This issue can be addressed by combining active learning—a machine learning form that determines which relation instances should be annotated for better extraction performance.

A large number of erroneous adverse drug events cannot be filtered out by co-occurrence analysis methods. The proposed model can more effectively extract

ADR signals from social media, significantly reducing the noise and redundancy of social media data and improving the accuracy of obtaining patient-reported adverse drug events.

5. Conclusion and Future Work

The emergence of social media provides a new approach for collecting healthcare data. Applying natural language processing techniques to extract patient-reported ADR signals from social media has great potential to improve the clinical and scientific knowledge of pharmacovigilance. However, extracting patient-reported ADR signals from social media remains a major challenge in medical informatics research.

This study developed a research framework using advanced natural language processing techniques to extract patient-reported ADR signals. The framework includes three main components: medical entity recognition for discussed drugs and events, adverse drug event relation extraction, and report source classification. Medical entity recognition employs a multi-dictionary source approach to address the diversity and colloquialism of online language expressions. The shortest dependency path kernel-based statistical learning method extracts entity relations, followed by semantic filtering based on medical knowledge and rules to further improve the accuracy of adverse drug event relation extraction. Finally, report source classification distinguishes actual patient-reported adverse drug events from rumors. To evaluate the proposed framework's performance, data collected from diabetes forums validated its effectiveness, with results showing that each component contributes to overall performance improvement. Future research directions include applying the framework to analyze treatments for different diseases and extract patient-reported content from other related social media topics.

References

[1] Liang Shaoxing, Li Fenglin. Research Status and Prospects of Context-Aware Health Information Service Systems[J]. Chinese Journal of Medical Library and Information Science, 2014(7): 31-36.

[2] Wang Dan. Active Monitoring of Adverse Drug Reactions and Its Development Trends[J]. Chinese Journal of Pharmacovigilance, 2015(10): 600-602.

[3] Harpaz R, DuMouchel W, Shah N, et al. Novel data-mining methodologies for adverse drug event discovery and analysis[J]. Clinical pharmacology and therapeutics, 2012, 91(6): 1010-1021.

[4] Leaman R, Wojtulewicz L, Sullivan R, et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks[C]//Proceedings of the 2010 workshop on biomedical natural language processing. Berlin: Association for Computational Linguistics, 2010: 117-125.

[5] Bian J, Topaloglu U, Yu F. Towards large-scale Twitter mining for drug-related adverse event[C]//Proceedings of the 2012 international workshop on smart health and wellbeing. New York: ACM, 2012: 25-32.

[6] Chee B, Berlin R, Schatz B. Predicting adverse drug events from personal health messages[C]//AMIA annual symposium proceedings. Bethesda: American Medical Informatics Association, 2011: 217.

[7] Wang Liwei. Research on Integration and Data Mining of Adverse Drug Event Information Resources[D]. Changchun: Jilin University, 2014.

[8] Abed S, Rachel G, Azadeh N, et al. Utilizing social media data for pharmacovigilance: a review[J]. Journal of biomedical informatics, 2015, 54(C): 202-212.

[9] Thomas P, Neves M, Solt I, et al. Relation extraction for drug-drug interactions using ensemble learning[C]//Proceeding of the 1st challenge task on drug-drug interaction extraction. Huelva: Tranging, 2011: 11-18.

[10] Liu X, Chen H. A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports[J]. Journal of biomedical informatics, 2015, 58: 268-279.

[11] Benton A, Ungar L, Hill S, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation[J]. Journal of biomedical informatics, 2011, 44(6): 989-996.

[12] Yang C, Jiang L, Zhang M. Social media mining for drug safety signal detection[C]//Proceedings of the 2012 international workshop on smart health and wellbeing. New York: ACM, 2012: 33-40.

[13] Dai Fei, Chen Shengxin, Shu Lixin, et al. Analysis and Application of Five Signal Mining Methods in Adverse Drug Reaction Detection[J]. Chinese Journal of Hospital Pharmacy, 2012, 32(20): 1674-1677.

[14] Bunescu R, Mooney R. A shortest path dependency kernel for relation extraction[C]//Proceedings of the conference on human language technology and empirical methods in natural language processing. Vancouver: Association for Computational Linguistics, 2005: 724-731.

[15] Li J, Zhang Z, Li X. Kernel-based learning for biomedical relation extraction[J]. Journal of the American Society for Information Science and Technology, 2008, 59(5): 756-769.

[16] Vincze V, Szarvas G, Farkas R, et al. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes[J]. BMC bioinformatics, 2008, 9(S11): 1-9.

[17] Uzuner Ö, Goldstein I, Luo Y. Identifying patient smoking status from medical discharge records[J]. Journal of the American Medical Informatics Association, 2008, 15(1): 14-24.

[18] Liu X, Chen H. AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums[C]//Proceedings of the international conference on smart health (ICSH 2013). Berlin: Springer, 2013: 134-150.

[19] Joachims T. Transductive inference for text classification using support vector machines[C]//Sixteenth international conference on machine learning. San Francisco: Morgan Kaufmann, 1999: 200-209.

Author Contributions

Wei Wei: Proposed research ideas, organized and wrote the paper.
Zheng Du: Implemented experiments, collected data, and analyzed results.

Note: Figure translations are in progress. See original paper for figures.