Evaluation of Text-Processing Algorithms for Adverse Drug Event Extraction from Social Media

Alejandro Metke-Jimenez
The Australian e-Health Research Centre
CSIRO, Australia
alejandro.metke@csiro.au

Sarvnaz Karimi Cecile Paris
Computational Informatics
CSIRO, Australia
sarvnaz.karimi, cecile.paris@csiro.au

ABSTRACT

The discovery of suspected adverse drug reactions is no longer restricted to mining reports that pharmaceutical companies and health professionals send to regulators for possible safety signals. Patient forums and other social media are being studied for additional sources of information to assist in expediting adverse reaction discovery. Extracting information on drugs, adverse drug reactions, diseases and symptoms, or patient demographics from such media is an essential step of this process, but it is not straightforward. While most studies in this area use a lexicon-based information extraction methodology, they do not explicitly evaluate the impact of text-processing steps on their final results. We experimentally quantify the value of the most popular techniques to establish whether or not they benefit the information extraction process.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis; J.3 [**Life and Medical Sciences**]: Health

Keywords

Adverse drug reaction discovery; information extraction; social media; text processing.

1. INTRODUCTION

An *Adverse Drug Event* (ADE) is an injury caused by a medication. This injury can be an unintended consequence of the drug's recommended usage, a consequence of its off-label usage, or a medication error. Adverse Drug Reactions (ADRs) are a subset of ADEs representing injuries caused by a drug administered at the recommended dosage for recommended symptoms. ADRs, also known as drug side effects, are a major concern for public health, costing millions of dollars worldwide to the healthcare systems [5, 7, 13]. The cost and limitations of traditional pharmacovigilance methods have prompted the need to explore additional sources that can be useful in the identification of potential *signals* of adverse drug reactions, which can then be used to conduct more thorough reviews. These reviews, performed by experts in regulatory agencies such as the Food and Drug Administration (FDA), intend to establish the causal relationship between an observed ADE and a drug. If such causality is confirmed and it is established that the drug is used as prescribed, further action is taken depending on the severity of the ADR. If the ADR is life-threatening, the drug is withdrawn from the market; otherwise the ADR is added to the drug's list of potential side effects.

Social media is one of the sources that potentially carries first-hand information that could be relevant for the ADR signal detection. A Pew survey[1] in 2009 [6] showed that 61% of American adults looked for health information online, 41% had read about someone else's experience, and 30% were actively creating new content.

Several attempts at extracting ADR signals from social media have shown promising results [2, 10, 15, 12], but one of the key steps in the signal detection process, identifying terms related to adverse events in noisy text, has not (to our knowledge) been studied on its own. This step is critical in the signal detection process, and errors can affect the subsequent stages of the process. We evaluate key proposed algorithms and some additional variations against a manually annotated data set of medical forum posts from the AskaPatient[2] medical forum.

Our contributions are threefold:

- measuring the value of different tokenisation algorithms, stemming, and stopping (i.e., stopword removal) for extraction of ADR-related terminology;
- measuring the effect of using a medical controlled vocabulary to filter out non-medical terms; and,
- measuring the value of using a consumer controlled vocabulary to filter out non-medical terms.

2. BACKGROUND

Extracting reports of ADEs from social media, in particular patient forums, has been studied since 2010. Although there is a large body of literature on information extraction from social media, especially Twitter, there is limited work on the specific area of ADE detection from social media. In

SoMeRA'14, July 11, 2014, Gold Coast, Queensland, Australia.
Copyright 2014 ACM 10.1145/2632188.2632200/14/07 ...$15.00.
http://dx.doi.org/10.1145/2632188.2632200.

[1] Pew surveys are public opinion surveys conducted by The Pew Research Center's Global Attitudes Project on a broad range of subjects.

[2] http://www.askapatient.com/

this section we first provide some background on controlled vocabularies that are commonly used in the relevant literature, then review the most relevant studies, and finally we explain how our work supplements their contributions.

2.1 Domain-Specific Vocabularies

Most previous studies employed one of the four controlled vocabularies below:

CHV. The Consumer Health Vocabulary (CHV)[3] provides a list of health terms used by lay people. It also contains frequent misspellings used by non-professionals. For example, it links both *lung tumor* and *lung tumour* to *lung neoplasms.* CHV is considered a promising candidate to assist in extracting health-related information from social media.

MedDRA. The Medical Dictionary for Regulatory Activities (MedDRA) is a thesaurus of ADRs used internationally by regulatory agencies and pharmaceutical companies to consistently code ADR reports [1]. Before MedDRA, the FDA had developed the Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) which is now obsolete.

SIDER. The public Side Effect Resource (SIDER) [9] contains a list of drugs and their known side effects as reported in different resources such as the FDA reports.

SNOMED CT. SNOMED Clinical Terms is a large ontology of medical concepts that has been recommended as the reference terminology for clinical information systems in countries such as Australia, the United Kingdom, Canada, and the United States [11]. It includes formal definitions, codes, terms, and synonyms for more than 300,000 medical concepts.

UMLS. The Unified Medical Language System (UMLS)[4] is a collection of several health and biomedical controlled vocabularies, including MedDRA, SNOMED CT, and CHV. UMLS provides a mapping between the concepts from each source. Also, it provides a semantic network that contains semantic types linked to each other through semantic relationships. Each concept is assigned one or more semantic types.

2.2 ADE Extraction from Social Media

Medical forums are online sites where people discuss their health concerns and share their experience with other patients or health professionals. Actively mining these forums could potentially reveal safety concerns regarding medications before regulators discover them through more passive methods via official channels such as health professionals.

We list some of the relevant studies that used social media data to extract ADEs in Table 1. Leaman et al. [10] proposed to mine patients' comments on health related web sites, specifically DailyStrength[5], to find mentions of adverse drug events. They used a lexicon that combines COSTART and a few other sources to extract ADR-related information from text. In a preprocessing step, they break the posts into sentences, tokenise the sentences, run a Part-of-Speech (POS) tagger, remove stopwords, and run the Porter stemmer. Using a sliding window approach, they match the lexicon entries with the preprocessed text and then evaluate the matches against the manually annotated text. Their data was annotated for adverse effect (same as ADE), beneficial effect, indication, and other. The paper did not include an explanation on how POS tags were utilised, how accurately it worked on informal text such as forum posts, and what was the effect of stopping and stemming.

Chee et al. [4] (second entry, Table 1) applied classifiers to identify drugs that have the potential for becoming part of the watchlist of the US regulatory, the FDA. They used patients posts on Health and Wellness Yahoo! Groups. The text was processed to generate features for the classifiers. It is unclear whether stemming or stopping was performed on the text before the words from the posts were extracted as features. Authors however mentioned that they did not fix any misspellings. They had two sets of features: all the words from the posts, and only those words that match their controlled vocabulary which included MedDRA and a list of diseases.

Benton et al. [2] extracted potential ADEs from a number of different breast cancer forums (such as `breastcancer.org`) using frequency counts of terms in a lexicon, controlled vocabulary in their corpus and then using association rule mining to establish the relationship between the matching terms. Association rule mining is a data-mining approach popular for mining ADEs from regulatory and administrative databases. The method by Benton et al. was an advancement on Leaman et al.'s [10] approach, as they did not stop at just the extraction of interesting concepts, but also proposed a method to establish a relationship between the extracted terms. However, again they performed no evaluation of their preprocessing step and how it affected their results.

Yang et al. [15] studied signal detection from a medical forum called MedHelp using data mining approaches. They extended the existing association rule mining algorithms by adding "interestingness" and "impressiveness" metrics. To process the forum data and calculate confidence and leverage, they had to find mentions of ADEs in the text. To do this, they used a sliding window and the CHV as the controlled vocabulary to match the terms.

Liu et al. [12] implemented a system called AZDrugMiner. Data was collected using a crawler and therefore required cleaning of the HTML tags and extracting of the text for further analysis. They then used a natural language processing tool called OpenNLP to break the text into sentences. To find relevant parts of each sentence, for example, mentions of a drug, they used MetaMap which maps text to UMLS concepts. After this stage, they extracted relations using co-occurrence analysis. They also used a tool called NegEx [3] to identify negations in the text. To our understanding, there was no stopping or stemming involved in the text processing of this study.

None of the studies mentioned above evaluated the effect of the preprocessing steps on their own which we do in our experiments.

3. METHODOLOGY

We implemented a simple, lexicon-based term identification mechanism, similar to the ones described in the existing literature, and tested different combinations of preprocess-

[3] `http://www.consumerhealthvocab.org/`

[4] `http://www.nlm.nih.gov/research/umls`

[5] `http://www.dailystrength.org/`

Table 1: Specifications of different ADR extraction studies on social media.

Study	Data	Controlled Vocab.	Preprocessing
Leaman et al. [10]	DailyStrength	COSTART, SIDER, MedEffect, UMLS, a manually compiled set of colloquial terms	Sentence boundary detection, Tokenisation, POS tagging, Stopping, Stemming
Chee et al. [4]	Health & Wellness Yahoo! Groups	MedDRA, lists of diseases	Unknown
Benton et al. [2]	Breast cancer forums	CHV, Cerner Multum's Drug Lexicon, Dietary supplements manually compiled, list of adverse events from AERS database	Stemming
Yang et al. [15]	MedHelp	CHV, a set of manually compiled terms	Remove punctuations, Stopping, Stemming
Liu et al. [12]	Diabetes online community	UMLS, MedDRA, and CHV	Text cleaning by removing HTML tags, removing URLs, removing punctuations, Sentence boundary detection

ing techniques and controlled vocabularies. The following sections describe these variables in detail. Note that we did not evaluate the use of complex natural language processing tools such as MetaMap.

3.1 Preprocessing

Before the text in the forum posts is matched against a controlled vocabulary it needs to be preprocessed. This involves two major steps: tokenisation and filtering. In the tokenisation step, the raw text is split into tokens. Three tokenisers were used:

- a simple *whitespace* tokeniser that splits the text on white spaces;
- a simple *letter* tokeniser that splits the text on non-letter characters;
- a *grammar-based* tokeniser that splits the text according to the Unicode Text Segmentation algorithm.[6]

In the filtering step, tokens generated in the previous step can be discarded or modified. The following filters were used:

- tokens that are stopwords[7] were either kept or removed.
- tokens were either stemmed using the well-known Porter stemmer or left as is.
- tokens were always transformed into lower case.

3.2 Matching

Once the text has been transformed into tokens, a controlled vocabulary is used to identify text fragments that refer to relevant concepts. The key variables in this step are the controlled vocabulary and the type of entity that is being identified. The following controlled vocabularies were used:

CHV As mentioned before, the main advantage of this source is that it contains health expressions used by consumers, not professionals; therefore, intuitively, it appears to be the best choice when dealing with social media.

UMLS UMLS includes terms from many controlled vocabularies including CHV, SNOMED CT, and MedDRA. In our experiments, we focus on the sources that come from technical medical terminologies; so we exclude entries originating from CHV.

ALL We also used the full 2013 active release of UMLS without any filtering, including the terms originating from CHV.

Table 2: Types of entities and the corresponding UMLS semantic types.

Entity Type	UMLS Semantic Type
Adverse drug reaction	Sign or symptom
Disease	Disease or syndrome
Medication	Organic chemical Pharmacological substance Clinical drug

Other sources used in related work include terms obtained from web scraping of health related web sites and resources created manually. These approaches are difficult to replicate and have therefore been excluded from this paper. We also exclude MedDRA because it is freely available only to the regulatory agencies.

All of the controlled vocabularies used in our implementation are linked to UMLS. This has the advantage that semantic types can be used to filter the vocabularies depending on the type of entity that needs to be identified. Table 2 shows the type of entities considered in this paper and the semantic types in UMLS used to filter the controlled vocabularies.

4. EXPERIMENTS

This section describes the ground truth data and metrics used to evaluate the different extraction algorithms.

4.1 Dataset

The AskaPatient medical forum provided us with all the posts on drugs containing Diclofenac from 2001 till 2013. It consisted of 250 posts and their responses, related to the drugs Arthrotec, Cambia, Cataflam, Diclofenac potassium, Diclofenac sodium, Flector, Pennsaid, Solaraze, Voltaren, and Zipsor. These posts were annotated by a group of four medical students using the Brat annotation tool[8]. The students were asked to annotate any symptom, medication and adverse drug reaction they could identify in the text. Annotation guidelines were similar but simplified as compared to the ones in [8]. Table 3 shows the list of tags that were available to the annotators.

The documents were divided evenly between the four annotators, except for five documents that were given to all the annotators for the purpose of calculating the inter-annotator agreement. Two metrics were used for this calculation: strict agreement and relaxed agreement. Both of these metrics are defined as the average of the pair wise agreement between the annotators. The agreement between each pair of anno-

[6] `http://unicode.org/reports/tr29/`

[7] http://norm.al/2009/04/14/list-of-english-stop-words/

[8] http://brat.nlplab.org/

Table 3: Tags and their definitions for the annotation.

Tag	Definition	Example
Drug	Mentions of the name of a medicine or drug	Diclofenac
ADR	Mentions of adverse drug reactions	Dizziness
Disease	Name of a disease for which the patient takes the medicine	Anxiety
Symptom	Symptoms of a disease that led them taking to a drug	My heart was racing

Table 4: Average pair-wise agreement between annotators.

Span	Annotation	Agreement
Strict	Strict	0.466
Strict	Relaxed	0.491
Relaxed	Strict	0.687
Relaxed	Relaxed	0.779

tators is defined as:

$$agreement(A_i, A_j) = \frac{max(count(A_i), count(A_j))}{count(match(A_i, A_j, \alpha, \beta))},$$

where A_i represents the annotations by the first annotator, A_j represents the annotations by the second annotator, and $match(A_i, A_j, \alpha, \beta)$ is a function that counts the number of matching tags.

The *match* function has two parameters. The first one, α, is the strictness of the spans. If span matching is configured to be strict, then the annotations being compared must match exactly. Consider the sentence "I experienced increased muscle tension". If annotator A_i annotates the text fragment "muscle tension" and annotator A_j annotates the text fragment "increased muscle tension", then the *match* function with strict span matching will return no matches. If span matching is configured to be relaxed, then the annotations that overlap will be counted as a match, with the restriction that each annotation can only be matched to one other annotation. In the previous example, the function would count the spans as a match. If annotator A_j had annotated the text fragments "increased" and "tension" instead, the function would still only count one match, because the fragment "increased muscle tension" would be mapped to the first overlapping span found, which in this case is "increased.

The second parameter, β, is the strictness in the *content* of the annotations. The content can be taken into consideration (strict) or ignored (relaxed). For example, if both annotators annotate the same text fragment, for example "muscle tension", but one of them uses the tag ADR, and the other one uses the tag Symptom, then the function will return a valid match only if the strictness of content parameter is relaxed. Table 4 shows the inter-annotator agreement using different configurations of the agreement metric. When span and annotation settings were both relaxed, the average agreement was approximately 78%. On average, each document contained approximately 5 annotations.

4.2 Evaluation Metrics

Similar to the calculation of inter-annotator agreement, the evaluation of the algorithm can be strict or relaxed. Strict evaluation requires that the spans match exactly. Spans are calculated based on character offsets as generated by our annotation tool Brat. Relaxed evaluation will match two overlapping spans, but a resulting span can only be mapped to a single ground truth span. However, if more than one resulting span completely covers a ground truth span then it is also considered a match. Consider the following example:

- ground truth = (182, 197) "aches and pains"
- actual result = (182, 187) "aches", (192, 197) "pains"

In strict evaluation mode, none of the resulting spans match and therefore both are considered false results. However, notice that the combination of both resulting tags matches the ground truth tag. In relaxed evaluation mode this is considered a single match.

Similar tasks have been evaluated in previous research. For example, Task 1A in the ShARe/CLEFeHealth2013 evaluation lab [14] evaluated the correctness in identification of spans of disorders in clinical reports using precision, recall, and F-Score. These metrics are defined as:

$$\text{Precision} = \frac{n_{TP}}{n_{TP} + n_{FP}},$$

$$\text{Recall} = \frac{n_{TP}}{n_{TP} + n_{FN}},$$

$$\text{F-Score} = 2 * \frac{precision * recall}{precision + recall},$$

where n_{TP} is the number of matching spans, n_{FP} is the number of spans reported by the system that are not part of the ground truth, and n_{FN} is the number of spans in the ground truth that were not reported by the system. We use these metrics in our evaluations.

4.3 Implementation

The system was implemented using Lucene[9] and consists of two main parts: the indexer and the searcher. The indexer is used to create an inverted index of the AskaPatient posts using the different combinations of tokenisers and filters mentioned in the previous section. The inverted index also stores additional information that is required to identify the positions of the terms in the documents (i.e., posts).

The searcher is used to import the controlled vocabularies and use their entries as queries. Each controlled vocabulary is imported from its source files, and, for each concept label a phrase query is issued against the index. The same tokenisers and filters used to create the index are used to process the queries. The results are then processed to extract the matching documents and the positions of the phrases in them. The system also merges overlapping spans created from synonyms of the same concept. Each query returns a set of documents and spans.

The final result is assembled by merging the results of all the queries. The system is then able to compare these results with the results from the ground truth.

4.4 Results and Discussion

[9] `https://lucene.apache.org/`

Table 5: Top F-Scores for ADR identification.

Evaluation	Tokeniser	Stopping	Stemming	Vocabulary	Precision	Recall	F-Score
Strict	Letter	No	No	CHV	0.463	0.438	**0.450**
	Grammar	No	No	CHV	0.463	0.438	**0.450**
	Grammar	No	Yes	CHV	0.438	0.454	0.446
	Letter	No	Yes	CHV	0.438	0.454	0.446
	Letter	Yes	No	CHV	0.453	0.429	0.441
Relaxed	Letter	No	No	CHV	0.709	0.649	**0.678**
	Grammar	No	No	CHV	0.708	0.649	0.677
	Letter	No	Yes	CHV	0.671	0.673	0.672
	Grammar	No	Yes	CHV	0.671	0.672	0.672
	Letter	Yes	No	CHV	0.701	0.644	0.672

Table 6: Top F-Scores for disease identification.

Evaluation	Tokeniser	Stopping	Stemming	Vocabulary	Precision	Recall	F-Score
Strict	Grammar	Yes	No	CHV	0.142	0.654	**0.233**
	Letter	Yes	No	CHV	0.094	0.654	0.165
	Grammar	Yes	Yes	CHV	0.085	0.673	0.151
	Grammar	Yes	No	UMLS	0.079	0.654	0.141
	Grammar	Yes	No	ALL	0.079	0.654	0.141
Relaxed	Grammar	Yes	No	CHV	0.161	0.745	**0.265**
	Letter	Yes	No	CHV	0.108	0.745	0.188
	Grammar	Yes	Yes	CHV	0.099	0.768	0.176
	Grammar	Yes	No	UMLS	0.090	0.745	0.160
	Grammar	Yes	No	ALL	0.090	0.745	0.160

The top 5 results for ADR identification for each evaluation type are shown in Table 5. The best F-Score for the strict evaluation is 0.45 (precision = 0.46 and recall = 0.44), while the best score for the relaxed evaluation is 0.67 (precision = 0.71 and recall = 0.65). The results show that the best combination for both the strict and relaxed evaluations is the letter tokeniser with no stopword removal and no stemming, followed closely by the grammar tokeniser also using no stop word removal and no stemming. The controlled vocabulary that produces the best results is CHV.

The top 5 results for disease identification for each evaluation type are shown in Table 6. The best F-Score for the strict evaluation is 0.23 (precision = 0.14 and recall = 0.65), while the best score for the relaxed evaluation is 0.27 (precision = 0.16 and recall = 0.75). The results show that the best combination for both the strict and relaxed evaluations is the grammar tokeniser with stopword removal and no stemming. The controlled vocabulary that produces the best results in this case is also CHV.

Notice that, in this case, the precision drops considerably compared to the results observed for ADR identification. One possible reason for this is that the annotation guidelines used to produce the ground truth instruct the annotators to only annotate diseases that the patient is experiencing. For example, in the sentence "After 3 years of having Ativan keep the anxiety & aggression in check", both "anxiety" and "aggression" will be annotated. On the other hand, in the sentence "Benadryl is an antihistamine and antihistamines are used for allergic reactions such as hayfever, hives, itching, runny nose. However, because...", it is not clear if the person is taking the medicine for "hayfever" or some other problem. Therefore these diseases will not be annotated. The algorithms evaluated in this paper will not be able to identify this difference, and hence this might explain the drop in precision.

Finally, the top 5 results for drug identification for each evaluation type are shown in Table 7. The best F-Score for the strict evaluation is 0.38 (precision = 0.25 and recall = 0.77), while the best score for the relaxed evaluation is 0.42 (precision = 0.28 and recall = 0.84). The results show that the best combination for both the strict and relaxed evaluations is the grammar tokeniser with stopword removal and no stemming. The controlled vocabulary that produces the best results in this case is also CHV.

In this case the precision is also quite low, but, unlike disease identification, the annotation guidelines indicate that all drugs in the posts should be annotated. Therefore, the reason for the drop in precision must lie elsewhere. After inspecting a subset of the automatically generated mappings, it is clear that one of the reasons for the low precision is that some text fragments in the posts generate multiple matches in the controlled vocabulary. The system was implemented with the ability of collapsing multiple spans into a single span when more than one synonym of a concept is mapped to the same span. However, the system keeps multiple overlapping spans if these belong to different concepts. The controlled vocabularies used for drug identification contain several overlapping descriptions for different concepts. For example, the concepts "Naproxen" and "Naproxen sodium" will create overlapping spans when the text "naproxen" is found in a post. Note that these are two separate concepts and are both included in the controlled vocabulary because both have the same semantic types (Organic Chemical and Pharmacologic Substance).

In these cases it is possible to select a single annotation, provided that the system has access to an underlying taxonomy or ontology. For example, UMLS contains a subclass relationship between "Naproxen" and "Naproxen sodium", so it is possible to determine that "Naproxen sodium" is more specific. This information could be used to eliminate overlapping annotations when one is found to be more specific than the others and this could improve the precision. This is left as future work.

The results show that, overall, the grammar tokeniser performs best, followed by the letter tokeniser, although the difference in performance is negligible. The effect of removing or keeping stopwords seems to depend on the type of entity being identified. For diseases and drugs, removing

Table 7: Top F-Scores for drug identification.

Evaluation	Tokeniser	Stopping	Stemming	Vocabulary	Precision	Recall	F-Score
Strict	Grammar	Yes	No	CHV	0.254	0.766	**0.382**
	Letter	Yes	No	CHV	0.218	0.771	0.339
	Whitespace	Yes	No	CHV	0.258	0.484	0.336
	Grammar	Yes	Yes	CHV	0.213	0.775	0.334
	Letter	Yes	Yes	CHV	0.183	0.780	0.296
Relaxed	Grammar	Yes	No	CHV	0.282	0.839	**0.422**
	Letter	Yes	No	CHV	0.245	0.853	0.381
	Whitespace	Yes	Yes	CHV	0.283	0.533	0.370
	Grammar	Yes	No	UMLS	0.236	0.844	0.369
	Letter	Yes	No	ALL	0.206	0.858	0.332

stopwords helps, while the opposite is true for ADR identification. Stemming, on the other hand, consistently deteriorated performance. Finally, the results show that CHV outperforms the other formal terminologies in the context of social media, as expected.

The main limitation of this work is that we deal specifically with medical forums, a specific type of social media. Other types of social media, such as Twitter, for example, will have different characteristics, and therefore these findings might not be applicable.

5. CONCLUSIONS

Active monitoring for drug adverse reactions extends the passive methods to new sources of information including social media. The transition to these new sources is necessary as the shortcomings of the current systems has led to large expenses imposed by adverse drug events on healthcare systems. Medical forums contain rich and first hand information from the consumers; however, they may contain noise and mining them for useful information can be challenging.

Recently, a number of studies have examined lexicon-based information extraction for identifying drugs and their adverse events from medical forums. Most of these studies do not evaluate the effect of preprocessing techniques or the choice of controlled vocabularies on their own. We address this by comparing different combinations of these settings on a single dataset. We showed that the grammar-based tokeniser performed best. Stopword removal was useful when extracting diseases and drugs, but not ADRs. Stemming did not seem to provide any benefit and was even detrimental in most cases. We also found that the CHV controlled vocabulary, although limited, has advantages over a comprehensive but formal vocabulary such as UMLS in the context of social media.

Acknowledgments. AskaPatient provided us with their data for research purposes only.

6. REFERENCES

[1] Introductory guide MedDRA version 17.0. Technical Report MSSO-DI-6003-17.0.0, The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), March 2014.

[2] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. Leonard, and J. Holmes. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics*, 44(6):989–996, 2011.

[3] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

[4] B. Chee, R. Berlin, and B. Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, pages 217–226, Washington, DC, 2011.

[5] J. Ehsani, T. Jackson, and S. Duckett. The incidence and cost of adverse events in victorian hospitals 2003-04. *The Medical Journal of Australia*, 184(11):551–555, 2006.

[6] S. Fox and S. Jones. The social life of health information. *Washington, DC: Pew Internet & American Life Project*, pages 2009–12, 2009.

[7] B. Hug, C. Keohane, D. Seger, C. Yoon, and D. Bates. The costs of adverse drug events in community hospitals. *Joint Commission Journal on Quality and Patient Safety*, 38(3):120–126, 2012.

[8] S. Karimi, S. Kim, and L. Cavedon. Drug side-effects: What do patient forums reveal? In *Proceedings of the 2nd International Workshop on Web Science and Information Exchange in the Medical Web*, pages 14–15, Glasgow, UK, 2011.

[9] M. Kuhn, M. Campillos, I. Letunic, L. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(1):Article 343, 2010.

[10] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards Internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden, 2010.

[11] D. Lee, R. Cornet, F. Lau, and N. De Keizer. A survey of SNOMED CT implementations. *Journal of Biomedical Informatics*, 46(1):87–96, 2013.

[12] X. Liu and H. Chen. AZDrugminer: An information extraction system for mining patient-reported adverse drug events in online patient forums. In *Proceedings of the 2013 International Conference on Smart Health*, pages 134–150, Beijing, China, 2013.

[13] E. Roughead and S. Semple. Medication safety in acute care in Australia: Where are we now? part 1: A review of the extent and causes of medication problems 2002-2008. *Australia and New Zealand Health Policy*, 6(1):18, 2009.

[14] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. 2013.

[15] C. Yang, L. Jiang, H. Yang, and X. Tang. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In *Proceedings of ACM SIGKDD Workshop on Health Informatics*, Beijing, China, 2012.