Extracting Adverse Drug Reactions from Social Media

Andrew Yates, Nazli Goharian, and Ophir Frieder
Information Retrieval Lab
Department of Computer Science
Georgetown University
{andrew, nazli, ophir}@ir.cs.georgetown.edu

Abstract

The potential benefits of mining social media to learn about adverse drug reactions (ADRs) are rapidly increasing with the increasing popularity of social media. Unknown ADRs have traditionally been discovered by expensive post-marketing trials, but recent work has suggested that some unknown ADRs may be discovered by analyzing social media. We propose three methods for extracting ADRs from forum posts and tweets, and compare our methods with several existing methods. Our methods outperform the existing methods in several scenarios; our filtering method achieves the highest F1 and precision on forum posts, and our CRF method achieves the highest precision on tweets. Furthermore, we address the difficulty of annotating social media on a large scale with an alternate evaluation scheme that takes advantage of the ADRs listed on drug labels. We investigate how well this alternate evaluation approximates a traditional evaluation using human annotations.

1 Introduction

The benefits of mining social media are rapidly increasing with social media's increasing popularity and the increasing amount of social media data available. According to the 2013 Health Online survey conducted by Pew[1], 59% of U.S. adults have "looked online for health information in the past year," which suggests that social media may also contain a wealth of health-related information. Extracting ADRs from social media is an important task because it can be used to augment clinical trials; in the case of recent drugs, it is possible for unknown ADRs to be discovered (i.e., to discover side effects that are not listed on the drug's label) (McNeil et al. 2012). Similarly, unknown interactions between drugs can be discovered (White et al. 2013).

In this work, we explore concept extraction methods for detecting ADRs in social media. Precision is particularly important for this task because any unknown ADRs discovered must go through a long clinical investigation to be validated. We focus on ADR extraction via dictionary-based methods in which an ADR dictionary is created from the terms in an ADR thesaurus. The thesaurus is necessary to treat two synonymous ADR terms or phrases, such as "hair loss" and "alopecia," as the same concept. While it is also possible to perform concept extraction without a dictionary, concept extraction methods are often dictionary-based because extracted ADR concepts must appear in some knowledge base regardless of whether the concept extraction is dictionary-based; otherwise, there is no way to determine the relationships between extracted ADRs and the ADRs listed on drug labels.

Dictionary-based ADR extraction requires two components: a dictionary-based concept extraction method and an ADR thesaurus to identify relationships between ADR terms. We use an existing ADR thesaurus (Yates and Goharian 2013) and propose three dictionary-based concept extraction methodologies. We compare these methods against several previously proposed methods (baselines) using both a forum corpus and a Twitter corpus. We find that our proposed methods substantially outperform the previously proposed methods in several different scenarios; when used to filter the output of other methods, our Multinomial NB and LLDA methods achieve the highest precision and F1 on the forum corpus, respectively; our CRF methods achieves the highest precision on the Twitter corpus.

Annotations indicating the concepts expressed in a document are traditionally used both to evaluate concept extraction methods and to train them. To alleviate the difficulty of obtaining annotations for a large social media corpus, we propose an alternate evaluation scheme for concept extraction in the ADR domain that uses the ADRs listed on drug labels to predict the ADRs that a method should extract. We explore both how well this "known ADR" evaluation approximates an evaluation using annotations and whether "known ADR" groundtruth can be used as training data for supervised methods.

Our contribution is an exploration of how well our proposed and existing concept extraction methods can be used to detect adverse drug reactions (ADRs) in social media (i.e., forum posts and tweets), including: *(i)* the proposal of three concept extraction methods; *(ii)* an evaluation of how well our proposed methods and existing baselines perform on social media; *(iii)* the proposal of an alternate evaluation scheme for the ADR domain that does not require annotations, an exploration of how well this alternate evaluation approximates an evaluation using annotations, and an exploration of how well this alternate evaluation scheme's ground

[1] *http://www.pewinternet.org/Reports/2013/Health-online.aspx*

truth can be used as training data for supervised methods.

2 Related Work

Researchers have studied concept extraction within the domains of biomedicine (Ben Abacha and Zweigenbaum 2011; Dinh and Tamine 2011; Aronson and Lang 2010) and electronic health records (Eriksson et al. 2013; Kang et al. 2012; Khare et al. 2012). Most concept extraction methods use a concept dictionary either to aid the extraction process or to map extracted concepts to known concepts. Methods that do not utilize a dictionary generally treat concepts as high-level categories (e.g., the "medical treatment" category) rather than concepts (e.g., a specific medical treatment).

MetaMap (Aronson and Lang 2010) is a concept extraction system commonly used in the biomedical domain. MetaMap maps text to concepts in the UMLS Metathesaurus (Bodenreider 2004) using a variety of information such as part-of-speech tags and generated token variants. We use MetaMap as one of our baselines by restricting the concepts it matches to those in our ADR thesaurus.

Ben Abacha and Zweigenbaum (2011) use a conditional random field (CRF) to classify phrases into the medical problem, treatment, and test categories. Chowdhury and Lavelli (2010) use a CRF with dictionary lookup features to identify when diseases are mentioned in academic paper abstracts; they classify phrases as mentioning a disease or not mentioning a disease, but they do not identify the disease concept being mentioned. Our CRF method (section 3.3) has some features in common with previously proposed CRFs, but we avoid using feature classes (e.g., orthographic features) that rely on properties of formal biomedical documents. Several methods use chunking to identify concept candidates. Rajagopal et al. (2013) identify noun and verb chunks and apply part-of-speech rules to each chunk to identify concepts; they determine the similarity between concepts by checking for shared terms and using a knowledge base. Brauer et al. (2010) use concept extraction to perform enterprise search; they compare noun phrases in documents against an enterprise ontology graph, which they use to create a document concept graph that serves as a query.

Zhou, Zhang, and Hu (2006) describe MaxMatcher, an approximate dictionary lookup method that has been used to identify biomedical concepts in academic papers (Chen et al. 2009; Dinh and Tamine 2011; Zhou, Zhang, and Hu 2006). MaxMatcher weights each term in a concept by the term's significance in respect to that concept. A score is assigned to each potential dictionary match based on the number, frequency, and the weight of the concept terms they include. MaxMatcher uses a series of rules to determine what text can be matched against a dictionary concept (e.g., a candidate must "end with a noun or a number").

Eriksson et al. (2013) describe the process of creating a Danish ADR dictionary and use it to find ADRs in health records that exactly match dictionary entries. Khare et al. (2012) find clinical concepts in clinical encounter forms using exact matching. They reduce the number of false positives by employing rules restricting the type of concepts that can be found in each type of field in the forms. Cohen (2005) uses exact matching to find biomedical concepts in biomedical research papers. Before checking the dictionary for exact matches, Cohen uses domain-specific rules to generate variants of the term being matched. These rules, which are specific to the notation used in gene and protein names, perform normalization such as removing spaces from a term and changing Arabic numerals to Roman numerals.

Some previous work has focused specifically on extracting ADRs from social media. Leaman et al. (2010) matched an ADR dictionary against terms appearing in a bag-of-words sliding window to extract ADRs from discussion forums; for example, the bag-of-words "achy legs restless" would match the "achy legs" and "restless legs" ADRs in the ADR thesaurus, causing both the "achy legs" and "restless legs" ADRs to be extracted. Our Sliding Window baseline is based on this method. Benton et al. (2011) used bag-of-words sliding windows to identify terms that were significantly more likely to occur together with a window than they were to occur separately. Yates and Goharian (2013) described ADRTrace, a system for finding ADRs by matching terms against an ADR thesaurus and matching terms against ADR patterns mined from drug reviews (e.g., "pain in my leg" is matched by the pattern "<X>in my <Y>"). We use ADRTrace as one of our baselines.

3 Methodology

We propose three methods for extracting ADR concepts from social media posts. Each method takes as input a training set of documents (forum posts or tweets) with ADR annotations and an ADR thesaurus; each method outputs the set of ADR concepts expressed in each document. If a method receives the document "after taking drug *X* I've noticed a loss of hair," for example, the method should extract the "hair loss" ADR.

3.1 Labeled LDA

Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is an unsupervised topic model used to associate documents with latent topics related to the terms observed in the documents. LDA chooses topics that best explain the observed terms; there is no guidance as to which topics or what type of topics should appear. Labeled LDA (LLDA) (Ramage et al. 2009) is a supervised extension to LDA that allows input documents to be labeled with topics. This allows the topics to be defined by a human rather than by LDA, which allows us to define LLDA topics as ADR concepts.

The intuition behind using LLDA for ADR extraction is that LLDA can identify terms that increase the likelihood of an ADR concept being present despite the fact that the terms are not associated in the ADR thesaurus with the ADR concept. LLDA's topics correspond to ADR concepts. For example, topic 1 might correspond to "weight gain," topic 2 might correspond to "carpal tunnel syndrome," and so on. A *no concept* topic is also included to indicate that a document contains no ADRs. This topic is intended to account for frequently occurring terms that should not be associated with an ADR topic. Each document is associated with the *no concept* topic in addition to any appropriate ADR topics (i.e., topics corresponding to the ADRs expressed in the document). The number of LLDA topics corresponds to the number of possible ADRs, plus one for the *no concept* topic. A subset of LLDA's documents are labeled and used as a training set. ADRs are extracted from the remaining documents by using LLDA to assign topics to documents that each correspond to an ADR concept or to the *no concept* topic. For each document, every ADR concept with a weight higher than the *no concept* topic in the document is extracted; multiple ADRs are extracted from a document if multiple ADR concepts have a higher weight than the *no concept* topic. No ADRs are extracted if the *no concept* topic is given the highest weight in the document.

3.2 Naïve Bayes

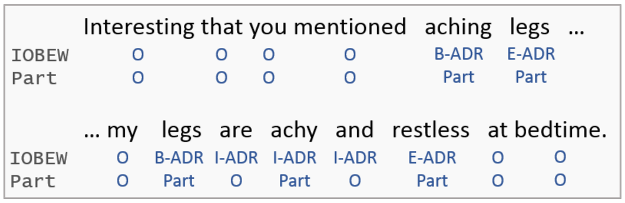
Naïve Bayes is a probabilistic classifier that performs document classification by computing the probability of a document belonging to a class given the document's terms; terms are assumed to be independent. This method can be used to perform ADR extraction by classifying documents as containing either "no concept" or an ADR concept (e.g., "hair loss"). We use Naïve Bayes with a multinomial model rather than a Bernoulli model, because this has been shown to have better performance on text classification tasks with a large vocabulary size (McCallum and Nigam 1998). Additive smoothing is used with $\alpha = 1$.

ADR mentions are often contained within small term windows, which are suitable for Naïve Bayes. Rather than using entire social media posts as documents, we treat each n-term sliding window in a social media post as a separate document. If a window contains an ADR, the window is labeled with that ADR concept; if not, the window will be labeled as *no concept*. When a window contains only part of an ADR phrase, it is labeled as *no concept*. This allows Naïve Bayes to learn that occurrence of part of an ADR phrase does not necessarily mean that the corresponding ADR concept should be extracted.

3.3 Conditional Random Field

A conditional random field (CRF) is a supervised discriminative probabilistic graphical model used for classification. Unlike most classifiers, CRFs consider the labels of each sample's neighbors when assigning a label to each sample. This makes CRFs well-suited to and commonly used for named entity recognition (NER) (Leaman and Gonzalez 2008). NER differs from concept extraction, however, in that its task is to identify categories (e.g., "Continent") rather than specific concepts within a category (e.g., "Antarctica"). NER labels often follow the IOBEW scheme (Leaman and Gonzalez 2008), which labels tokens with their relation to the entity (e.g., **I**nside or **O**utside the entity). Each entity must be either a single token (**W**) or have a **B**eginning and an **E**nd. We also use a simpler labeling scheme, Part (or "IO"), that labels only the tokens which are part of an ADR.

Figure 1 shows the labels assigned to two phrases using the IOBEW and Part schemes. The ADR "aching legs" in the top phrase is composed entirely of consecutive terms, so there is no difference between the schemes. The two ADRs in the bottom phrase (i.e., "achy legs" and "restless legs") are composed of non-consecutive terms; to handle these ADRs with the IOBEW scheme, the surrounding terms (i.e., "are" and "and") must also be labeled as part of the ADR entity. To extract ADRs with the IOBEW scheme, we treat the terms within each entity (i.e., between each *B-ADR* and *E-ADR*) as a bag of words and form all possible ADRs from those terms. To extract ADRs with the Part scheme, we treat all terms labeled *Part* as a bag of words and form all possible ADRs from those terms.



	Interesting	that	you	mentioned	aching	legs	...
IOBEW	O	O	O	O	B-ADR	E-ADR	
Part	O	O	O	O	Part	Part	

	... my	legs	are	achy	and	restless	at	bedtime.
IOBEW	O	B-ADR	I-ADR	I-ADR	I-ADR	E-ADR	O	O
Part	O	Part	O	Part	O	Part	O	O

Figure 1: CRF labels with the IOBEW and Part schemes.

We use a first-order CRF and associate the following Boolean features with each token: (1) token itself, (2), dependency relation types that the token appears in as determined by the Stanford Parser (Klein and Manning 2003; Marneffe, Maccartney, and Manning 2006), (3) token's part-of-speech tag, (4) token appearance anywhere in the ADR thesaurus, (5) tokens immediately before and after the current token, and (6) part-of-speech tags immediately before and after the current token. We do not use orthographic features (e.g., the capitalization patterns of tokens) that are commonly used in biomedical NER (Bundschus et al. 2008), as common language ADR expressions do not frequently follow any orthographic patterns. Such features are more appropriate for protein or gene names, which commonly have capital letters and hyphens. While some of our features are derived from part-of-speech tags and dependency relations, we do not rely on the presence of certain tags or relations to identify candidates. To perform part-of-speech tagging, we use the Stanford Parser on the forum corpus and the TwitIE tagger (Bontcheva et al. 2013) on the Twitter corpus.

3.4 Filtering

LLDA and Multinomial NB can be used as a filter for other methods to substantially increase their precision, because *(i)* they make different types of classification errors than the other methods and *(ii)* they have a low false negative rate. The CRFs and baseline methods do not meet these criteria, so their performance is poor when used as filters; for the sake of brevity, we only report results using LLDA and Multinomial NB filters. When one of these methods is used as a filter, only ADRs extracted by both a filtering method (i.e., LLDA or Multinomial NB) and another ADR extraction method are extracted (i.e., we return the intersection of the ADRs extracted by a filtering method and the ADRs extracted by another method). A filter cannot improve recall because it can only cause fewer results to be extracted, but a high recall filter can improve precision and F1 by reducing the number of false positives while introducing a minimal number of false negatives.

3.5 Baselines

We use ADRTrace (Yates and Goharian 2013), MaxMatcher (Zhou, Zhang, and Hu 2006), and a sliding window method based on the approach described by Leaman et al. (2010) as our baselines. These methods are described in the related work section (section 2).

4 Evaluation

We evaluate the methods' performance when used to extract ADRs from medical forum posts and tweets. We first describe the ADR thesaurus, forum post corpus, Twitter corpus, and groundtruth used to evaluate the methods. We then report results on the forum corpus and demonstrate that using a method as a filter can substantially improve the results. Next we evaluate the methods' performance using the alternate "known ADR" evaluation, which does not require human annotations as groundtruth, and compare the evaluation's results to those obtained using annotations as ground truth. We use the "known ADR" evaluation to evaluate the methods' performance on the Twitter corpus. Finally, we investigate how well the alternate evaluation's ground truth can be used to provide training data for the forum corpus.

4.1 Experimental setup

Thesaurus We use MedSyn (Yates and Goharian 2013), a thesaurus of ADR terms derived from a subset of the Unified Medical Language System Metathesaurus (Bodenreider 2004), as our list of synonyms for each ADR concept. Every ADR extracted by any of the methods is mapped to a synonym in MedSyn (e.g., "hair loss" is mapped to the ADR concept "alopecia") and then compared against our ground truth (sections 4.1 and 4.2). MedSyn includes both common language terms (e.g., "baldness," "hair loss") and expert medical terms (e.g., "alopecia").

Forum Corpus Our forum corpus consists of 400,000 forum posts crawled from the Breastcancer.org and FORCE discussion forums[2]. The crawl was targeted at retrieving posts from sub-forums related to the discussion of ADRs caused by breast cancer drugs.

The forum corpus' ground truth consists of a random subset of the corpus that was annotated to indicate the ADRs expressed in each post. Human annotators were instructed to read each post and indicate any first-hand ADR accounts; third person accounts, that is, a person talking about another person's ADR experience, were ignored to avoid counting the same ADR experience multiple times. The annotators were also instructed to ignore negated ADRs and uncertain ADRs. Each post was annotated by three separate annotators. The annotators annotated approximately 600 posts with a combined total of 2,100 annotations. The MedSyn thesaurus was used to treat different synonyms of the same ADR as equivalent. Fleiss' Kappa was calculated to be 0.37, indicating sufficient inter-rater reliability, given the difficulty of a many-category annotation task. When annotators disagreed over whether an ADR was expressed, we resolved the conflict by taking a majority vote. The annotations and a list of the crawled URLs are available on the authors' website [3].

Drug	Unknown adverse drug reactions
Docetaxel	Low hemoglobin, low potassium, metallic tastes
Tamoxifen	Elbow pain, hives, hypermenorrhoea, restless leg syndrome, tearfulness

Table 1: Unknown ADRs present in the forum annotations.

Twitter Corpus Our Twitter corpus consists of approximately 2.8 billion tweets collected from Twitter's public streaming API over the time periods May 2009 – October 2010 and September 2012 – May 2013. We used the SIDER 2 (Kuhn et al. 2010) database, which contains drugs and their known side effects, to keep only those tweets that contained the name of at least one drug, leaving us with approximately 2.4 million tweets. Drug terms in tweets were matched against SIDER 2 using an exact dictionary lookup. SIDER 2 differs from larger databases by providing structured information about each drug's ADRs. This allows the ADRs to be mapped directly to MedSyn without performing an intermediate matching step. Both the tweet ids and the tweet-drug mappings for the tweets in our corpus are available on the authors' website [4]. In January 2014, we queried the Twitter API for each of these remaining tweets to keep only English language tweets and tweets that still existed (i.e., were not marked as spam, deleted by the author, or made private by the author). This resulted in a corpus of approximately 329,000 tweets. Rather than using annotations with the Twitter corpus, we use an alternate evaluation methodology (section 4.2).

Unknown ADR Annotations Given our goal of accurately extracting ADRs so that unknown ADRs may be discovered, it is instructive to look for unknown ADRs in the annotations themselves. The unknown ADRs (i.e., ADRs that are not listed on the associated drug's label) that appear in the forum annotations are shown in Table 1.

Only ADRs that were annotated at least twice for a drug are shown. While these ADRs are not known to be caused by their associated drugs, some may be caused by a medical procedure or underlying condition that commonly accompanies the drug. This effect was partially compensated for by excluding unknown ADRs that were present for more than one of the five breast cancer drugs.

Further clinical investigation is required to determine if any of the unknown ADRs could be caused by their associated drug rather than by other factors. "Low hemoglobin" is likely to be a result of the patient's underlying condition (i.e., breast cancer) even though it only appears as an ADR for Docetaxel. "Low potassium," however, is not clearly associated with breast cancer. Similarly, "tearfulness" is likely related to the emotional burden caused by the patient's condition, whereas "restless leg syndrome" could potentially

[2] *http://breastcancer.org* and *http://facingourrisk.org*

[3] *http://ir.cs.georgetown.edu/data/aaai15/*

[4] *http://ir.cs.georgetown.edu/data/aaai15/*

	Method extracted the ADR *x*	Method did **not** extract ADR *x*
ADR *x* is listed on the given drug's label	True positive	False negative
ADR *x* is **not** listed on the given drug's label	False positive	True negative

Table 2: Known ADR scoring

be an unknown side effect. These uncertainties illustrate the difficulty of determining whether a potential unknown ADR is actually an unknown ADR caused by a drug. Clinical studies would be required to determine which ADRs are truly unknown ADRs.

4.2 "Known ADR" evaluation

While our discussion forum corpus was annotated by humans, annotating documents is a difficult, time consuming process, which makes annotating a large corpus impractical. Instead, we propose an alternate evaluation methodology that performs distant supervising by taking advantage of the ADRs listed on drug labels (i.e., the list of ADRs printed on the drug's packaging) and does not require annotations. We call this setup a "known ADR" evaluation. This is similar to the semantic bootstrapping approach used by Mintz et al. (2009) to train a relation classifier with unlabeled sentences containing a pair of known entities.

In a known ADR evaluation, each document is associated with the drugs mentioned in the document; documents that do not mention any drugs are ignored. We assume that the ADRs listed on the mentioned drugs' labels are the ADRs that should occur in the documents the vast majority of the time (i.e., most of the ADRs that people mention should be known ADRs; if a significant fraction of a drug's ADRs are unknown, the drug should not be on the market and thus should not appear in our data).

We use these drug-document associations to define true positives, true negatives, false positives, and false negatives as shown in Table 2. For example, a true positive occurs when an extracted ADR is listed on the drug label of a drug associated with the given document. The known ADR evaluation uses SIDER 2 to associate each drug with its ADRs; SIDER 2 contains structured ADR data that can be mapped directly to MedSyn using UMLS concept identifiers, so no separate matching step is necessary to map ADRs in the known ADR ground truth to ADRs in MedSyn.

Calculating precision is straightforward because we assume that only ADRs listed on a drug's label should be extracted. While there are circumstances in which it is incorrect to extract an ADR listed on a drug's label, such as when the ADR is mentioned in a third-person account, we assume for the purposes of evaluation that these cases are infrequent; we investigate the validity of this assumption in section 4.5.

False negatives frequently occur because it is unlikely for a social media document (forum post or tweet) to mention every ADR that a drug can cause (i.e., every ADR listed on a drug's label); in fact, Twitter's character limit makes this impossible for most drugs. Subsequently, accurately calculating recall is impossible as we have no way of knowing the number of ADRs listed on a drug's label that were actually mentioned in a document. This difficulty does not change the relative ranking of different methods' recall, hence we consider the relative rankings of the methods' recall scores rather than their absolute recall scores. We do not report F1 with known ADR evaluations, as we cannot accurately estimate absolute recall scores.

4.3 Drug-aware cross-validation

Five-fold cross-validation is used with all reported results. Both stratified cross-validation folds and random folds, which are commonly used in supervised learning, are undesirable for this problem; they would allow supervised methods to learn the mapping of certain drug names to certain ADRs, which conflicts with our goal of identifying "unknown ADRs" (i.e., a method trained on random folds might perform well only when extracting known ADRs).

To avoid this situation, we choose cross-validation folds such that there is minimal overlap between the drugs represented in each fold. For example, if fold #1 contains documents mentioning the drugs D1 and D2, and fold #2 contains documents mentioning the drug D3, no documents containing D1, D2, or D3 should be placed in the three remaining folds. The drugs in each fold are chosen such that the number of documents in each fold is as close as possible to the number of document in each other fold. In practice, there can be some small overlap between the drugs in each fold due to documents mentioning multiple drugs. The problem of assigning drugs to folds can be viewed as an instance of the bin packing problem, in which items of varying sizes must be packed into bins with limited capacities (Vazirani 2003). Folds are treated as bins with a maximum capacity of 20% of the total number of documents (corresponding to five folds). Each distinct drug is treated as an item, and the item's size is the number of documents corresponding to that drug. The bin packing problem is NP-hard, so we obtain an approximate solution using the first-fit algorithm (Vazirani 2003).

4.4 Forum Evaluation

We evaluate each method's performance in extracting ADRs on our forum corpus by using the annotations as ground truth. For each method we calculate precision, recall, and F1 by comparing the set of ADRs in the annotations to the set of ADRs extracted by that method. We use a sliding window size of 5 (n=5) with the Sliding Window method and with multinomial Naïve Bayes (Multinomial NB); we empirically chose n=5 for both methods.

The results are shown in the "no filter" column of Table 3. Sliding Window performs best as measured by F1 and Recall. The CRFs perform best in terms of precision; CRF-Part (i.e., the CRF with the Part labeling scheme) has an 8% higher F1 than CRF-IOBEW, suggesting that the surrounding terms commonly found in ADR concept expressions may be handled better by the Part labeling scheme. The CRFs have higher precision than the other methods, but lower recall. MetaMap, MaxMatcher, LLDA, and Multinomial NB perform substantially worse than the other meth-

	No filter			LLDA filter			Multinomial NB filter			Known ADR	
	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	Prec.	Rec. Rank (Rec.)
ADRTrace	0.44	0.35	0.59	0.58	0.63	0.54	0.57	0.59	0.55	0.48	4 (0.00886)
MaxMatcher	0.23	0.25	0.21	0.18	0.58	0.10	0.18	0.58	0.10	0.46	8 (0.00177)
Sliding Window	0.46	0.36	0.65	0.58	0.58	0.57	0.55	0.54	0.57	0.49	3 (0.01075)
MetaMap	0.25	0.29	0.21	0.24	0.48	0.16	0.28	0.48	0.19	0.52	7 (0.00235)
CRF-IOBEW	0.37	0.58	0.28	0.38	0.64	0.27	0.40	0.66	0.28	0.90	6 (0.00458)
CRF-Part	0.40	0.58	0.31	0.39	0.57	0.30	0.40	0.59	0.30	0.87	5 (0.00499)
LLDA	0.11	0.06	0.58	-	-	-	0.38	0.29	0.55	0.58	2 (0.04565)
Multinomial NB	0.27	0.18	0.54	0.38	0.29	0.55	-	-	-	0.62	1 (0.20169)

Table 3: Results on the forum corpus using human annotations with no filtering method, using annotations with LLDA as a filter, using annotations with Multinomial NB as a filter, and using known ADR groundtruth instead of human annotations. When used as filters, LLDA and Multinomial NB increase the precision of other methods.

ods. Both were designed to operate on biomedical documents (i.e., academic papers) rather than on social media.

MaxMatcher's rules choose poor match candidates when used to perform concept extraction on social media. This results in both low recall, as correct candidates are missed, and in low precision, as MaxMatcher tries to extract ADRs from partial matches against poor match candidates. Similarly, MetaMap's integrated part-of-speech tagging does not utilize a part-of-speech tagger trained on social media.

The forum evaluation results when using LLDA and Multinomial NB as filters are shown in the 2nd and 3rd columns of Table 3, respectively. ADRTrace and Sliding Window with an LLDA filter both yield an F1 26% higher than Sliding Window alone, which was the best performing method without a filter. Combining CRF-IOBEW with an LLDA filter slightly improves its F1 by increasing its precision by 10%; CRF-Part's F1 drops slightly. MetaMap's precision and MaxMatcher's precision both increase substantially, but at the expense of their recall scores, indicating that these methods are identifying some ADRs that LLDA and Multinomial NB miss. The Multinomial NB filter's maximum F1 is slightly lower than the LLDA filter's; however, it may be preferable in some situations because it is much more computationally efficient than LLDA.

The performance improvement caused by using LLDA or Multinomial NB as a filter comes from their different approach to extraction (i.e., they do not require every term in an ADR concept to be present) and their high recall (i.e., they do not eliminate many true positives). If a method's utility as a filter relied only on the method's recall, Sliding Window should perform better than both LLDA and Multinomial NB; this is not the case. Using Sliding Window as a filter with any of the remaining methods yields a F1 no higher than 0.40, however, Sliding Window's recall is 12% higher than LLDA's. In addition, Sliding Window's precision is substantially higher than both of the other methods'.

4.5 Known ADR evaluation on forums

We perform a known ADR evaluation on our forum corpus to explore how well the known ADR evaluation's results match the annotated evaluation on the forum corpus. We use human annotations to train the methods and the known ADR ground truth to evaluate the methods' performance; if the known ADR evaluation is valid, the metrics should be similar to the metrics obtained when using annotations as ground truth.

The results are shown in the "Known ADR" column of Table 3. We do not report F1 due to the difficulty of accurately calculating recall scores (as explained in section 4.2). Each method's precision is higher than in the forum annotation evaluation (section 4.4) because the methods' task has become simpler; for example, the known ADR evaluation does not make a distinction between first-hand and third person ADR mentions. LLDA and Multinomial NB achieve artificially high recall scores because they learn to frequently return the most common ADRs.

Given the difficulty of directly comparing F1 and recall scores between the human annotation evaluation and known ADR evaluation, we compare the correlations between the absolute precision scores and the correlations between the recall rankings. We use Pearson's r to compare the correlation between the known ADR evaluation's precision and the annotation evaluation's precision. We focus on the ranking of the recall scores and use Spearman's ρ to compare the correlation between the known ADR's recall rankings and the annotation evaluation's recall rankings. The correlation in precision between the evaluation setups is 0.64, and the correlation of the recall ranks is 0.79. Both of these coefficients indicate a strong correlation, suggesting that, while the known ADR evaluation does not yield the same precision and recall scores as the annotation evaluation, the known ADR evaluation can be used to evaluate the relative ranking of methods. Evaluating the relative ranking of methods can be useful in determining if a method performs well on different data sets (e.g., to determine if the method that performs best on forum posts also performs best on tweets).

4.6 Twitter evaluation

To investigate the difference between the methods' performance on forum posts and their performance on tweets, we evaluate each method's performance on our Twitter corpus using the known ADR evaluation described in section 4.2. Recall that the known ADR evaluation's results on forum posts were strongly correlated with the annotation evaluation's results on forum posts (section 4.5).

The results are shown in Table 4. As in the forum annota-

	No filter		LLDA filter	
	Prec.	R. Rank	Prec.	R. Rank
ADRTrace	0.23	4 (0.00047)	0.25	3 (0.00048)
MaxMatcher	0.23	5 (0.00015)	0.24	5 (0.00018)
S. Window	0.23	3 (0.00054)	0.25	2 (0.00055)
MetaMap	0.24	7 (0.00014)	0.24	5 (0.00018)
CRF-IOBEW	0.44	5 (0.00015)	0.44	4 (0.00019)
CRF-Part	0.48	8 (0.00013)	0.43	7 (0.00016)
LLDA	0.20	1 (0.13664)	-	-
Multi. NB	0.37	2 (0.00194)	0.24	1 (0.00134)

Table 4: Twitter results using known ADR groundtruth. The CRF methods perform well as on the forum corpus, but the LLDA filter does not.

	F1	Precision	Recall
CRF-IOBEW	0.34	0.43	0.28
CRF-Part	0.35	0.41	0.31
LLDA	0.10	0.06	0.38
Multi. NB	0.19	0.12	0.39

Table 5: Forum results when known ADR ground truth is used as training data for the supervised methods. The CRF methods' precision scores decrease, but remain higher than those of any unsupervised method.

tion evaluation (section 4.4), CRFs achieve the highest precision both by themselves and with a filter. LLDA achieves the highest recall, whereas Sliding Window had the highest recall in the forum annotation evaluation. MaxMatcher and MetaMap perform similarly poorly in both evaluations. LLDA and Multinomial NB's unrealistically high recalls are likely caused by the known ADR evaluation, rather than by the different corpus, as LLDA and Multinomial NB were observed to have the highest recalls when the known ADR evaluation was applied to the forums (section 4.5). LLDA and Multinomial NB are learning to frequently return common ADRs caused by many drugs.

As explained in the forum known ADR evaluation (section 4.5), we believe the known ADR evaluation is most useful for evaluating the relative ranking of methods. CRF-Part is the highest ranked method by precision but has low recall; the Multinomial NB and Sliding Window methods have the 2nd and 3rd highest recalls, respectively.

We use Spearman's rank correlation coefficient to compare the precision and recall rankings between the Twitter corpus and the annotation evaluation on the forum corpus. The rank correlation between precision scores is 0.62 and the rank correlation between recall scores is 0.65. Both correlation coefficients indicate a strong correlation between the methods' rankings when used on the forum corpus and used on the Twitter corpus, suggesting that the methods that work well on forum posts also work well on tweets.

Without being able to calculate F1 and absolute recall it is difficult to determine the LLDA filter's overall performance impact, however, it does not appear to improve performance on the Twitter corpus as it did on the forum corpus. In fact, LLDA's and Multinomial NB's behavior as filter methods differs from that observed in the annotation evaluation on the forum corpus (section 4.4), where LLDA and Multinomial NB increased the precision of most other methods when used as filters. When given known ADR training data, LLDA and Multinomial NB learn to predict the most common ADRs across drugs, which prevents them from being useful as filtering methods.

4.7 Known ADRs as training data

We introduced the known ADR evaluation setup as a strategy for dealing with the difficulty of annotating large amounts of documents. Annotations are not only required for evaluation, however; the supervised CRF, LLDA, and Multinomial NB methods require annotations as training data. In this section we investigate the utility of using the known ADR ground truth (i.e., the assumption that all ADRs on a drug label should be extracted) as training data for our supervised methods. We conduct the evaluation using the forum corpus with known ADR ground truth to train the methods and the forum annotation evaluation (section 4.4) to test the methods.

The results using the forum corpus with known ADR ground truth as training data are shown in Table 5. The supervised methods perform worse than ADRTrace and Sliding Window did in the annotation evaluation in terms of F1 and recall, which suggests that known ADR ground truth should not be used as training data for F1-oriented or recall-oriented scenarios. CRF's precision is still higher than that of any of the unsupervised methods, however, indicating that known ADR ground truth may still be useful for training in precision-oriented scenarios.

5 Conclusion

We proposed three methods for performing concept extraction and evaluated their performance in the domain of adverse drug reaction (ADR) extraction from social media. We found that in different scenarios our methods showed performance advantages over previously proposed methods. Our CRF method's high precision, which remains the highest of the methods evaluated, makes the method well-suited for precision-oriented ADR extraction applications, such as detecting unknown ADRs that are not listed on a drug's label. Furthermore, our LLDA and Multinomial NB methods can be used as filters to improve the precision of other concept extraction methods; when Multinomial NB was used on the forum corpus to filter our CRF's extractions, the combination led to the highest precision of any method we evaluated; when LLDA was used as a filter with the Sliding Window or ADRTrace methods, the combined methods achieved the highest F1s of any methods we evaluated.

6 Acknowledgements

This work was partially supported by the US National Science Foundation through grant CNS-1204347.

References

Aronson, A. R., and Lang, F.-M. 2010. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc* 17(3):229–236.

Ben Abacha, A., and Zweigenbaum, P. 2011. Medical entity recognition: A comparison of semantic and statistical methods. In *BioNLP '11*, number ii, 56–64.

Benton, A.; Ungar, L.; Hill, S.; Hennessy, S.; Mao, J.; Chung, A.; Leonard, C. E.; and Holmes, J. H. 2011. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *Journal of biomedical informatics* 44(6):989–96.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research* 32(Database issue):D267–70.

Bontcheva, K.; Derczynski, L.; Funk, A.; Greenwood, M.; Maynard, D.; and Aswani, N. 2013. Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of Recent Advances in Natural Language Processing (RANLP '13)*, 83–90.

Brauer, F.; Huber, M.; Hackenbroich, G.; Leser, U.; Naumann, F.; and Barczynski, W. M. 2010. Graph-based concept identification and disambiguation for enterprise search. In *WWW '10*, 171. ACM Press.

Bundschus, M.; Dejori, M.; Stetter, M.; Tresp, V.; and Kriegel, H.-P. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics* 9.

Chen, X.; Lu, C.; An, Y.; and Achananuparp, P. 2009. Probabilistic models for topic learning from images and captions in online biomedical literatures. In *CIKM '09*, 495.

Chowdhury, M. F. M., and Lavelli, A. 2010. Disease mention recognition with specific features. In *BioNLP '10*.

Cohen, A. M. 2005. Unsupervised gene / protein named entity normalization using automatically extracted dictionaries. In *ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, 17–24.

Dinh, D., and Tamine, L. 2011. Voting techniques for a multi-terminology based biomedical information retrieval. In *Proceedings of the 13th conference on Artificial intelligence in medicine (AIME '11)*.

Eriksson, R.; Jensen, P. B. d.; Frankild, S.; Jensen, L. J.; and Brunak, S. r. 2013. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association : JAMIA* 1–7.

Kang, N.; Afzal, Z.; Singh, B.; van Mulligen, E. M.; and Kors, J. a. 2012. Using an ensemble system to improve concept extraction from clinical records. *Journal of biomedical informatics* 45(3):423–8.

Khare, R.; An, Y.; Li, J.; Song, I.-Y.; and Hu, X. 2012. Exploiting semantic structure for mapping user-specified form terms to SNOMED CT concepts. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 285. ACM Press.

Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *ACL '03*, volume 1, 423–430. ACL.

Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; and Bork, P. 2010. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology* 6:343.

Leaman, R., and Gonzalez, G. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 663, 652–63.

Leaman, R.; Wojtulewicz, L.; Sullivan, R.; Skariah, A.; Yang, J.; and Gonzalez, G. 2010. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. In *BioNLP '10*, 117–125.

Marneffe, M.-c. D.; Maccartney, B.; and Manning, C. D. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC-06*, 449 – 454.

McCallum, A., and Nigam, K. 1998. A comparison of event models for naive bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*.

McNeil, J. J.; Piccenna, L.; Ronaldson, K.; and Ioannides-Demos, L. L. 2012. The Value of Patient-Centred Registries in Phase IV Drug Surveillance. *Pharmaceutical Medicine* 24(5):281–288.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, ACL '09, 1003–1011. ACL.

Rajagopal, D.; Cambria, E.; Olsher, D.; and Kwok, K. 2013. A Graph-Based Approach to Commonsense Concept Extraction and Semantic Similarity Detection. In *Proceedings of the 22nd international conference on World Wide Web companion*, 565–570.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled LDA : A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP '09*.

Vazirani, V. V. 2003. *Approximation Algorithms*. Springer Berlin / Heidelberg.

White, R. W.; Tatonetti, N. P.; Shah, N. H.; Altman, R. B.; and Horvitz, E. 2013. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association (JAMIA)* 20(3):404–8.

Yates, A., and Goharian, N. 2013. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In *Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13)*.

Zhou, X.; Zhang, X.; and Hu, X. 2006. MaxMatcher: Biological concept extraction using approximate dictionary lookup. *PRICAI 2006: Trends in Artificial Intelligence* 4099:1145–1149.