Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark

Rachel Ginn¹, Pranoti Pimpalkhute¹, Azadeh Nikfarjam¹, MS, Apurv Patki¹, Karen O’Connor¹, Abeed Sarker¹, PhD, Karen Smith², PhD, Graciela Gonzalez², PhD

¹Arizona State University, Tempe, AZ; ²Regis University, Denver, CO
E-mail: rginn@asu.edu, ppimpalk@asu.edu, anikfarj@asu.edu, apatki1@asu.edu, Karen.Oconnor@asu.edu, abeed.sarker@asu.edu, ksmith003@regis.edu, Graciela.Gonzalez@asu.edu

Abstract
With many adults using social media to discuss health information, researchers have begun diving into this resource to monitor or detect health conditions on a population level. Twitter, specifically, has flourished to several million users and could present a rich information source for the detection of serious medical conditions, like adverse drug reactions (ADRs). However, Twitter also presents unique challenges due to brevity, lack of structure, and informal language. We present a freely available, manually annotated corpus of 10,822 tweets, which can be used to train automated tools to mine Twitter for ADRs. We collected tweets utilizing drug names as keywords, but expanding them by applying an algorithm to generate misspelled versions of the drug names for maximum coverage. We annotated each tweet for the presence of a mention of an ADR, and for those that had one, annotated the mention (including span and UMLS IDs of the ADRs). Our inter-annotator agreement for the binary classification had a Kappa value of 0.69, which may be considered substantial (Viera & Garrett, 2005). We evaluated the utility of the corpus by training two classes of machine learning algorithms: Naïve Bayes and Support Vector Machines. The results we present validate the usefulness of the corpus for automated mining tasks. The classification corpus is available from http://diego.asu.edu/downloads.

Keywords: adverse drug reactions, twitter, social media, mining, machine learning, biomedicine, pharmacovigilance, classification, natural language processing

1. Introduction

Adverse drug reactions (ADRs), defined as “injuries resulting from medical drug use”, present a significant health problem. A systematic review of twenty-five prospective observational studies determined that 5.3% of all hospital admissions are associated with adverse drug reactions (Kongkaew, Noyce, & Ashcroft, 2008). Accelerating the detection of these events could greatly impact human health.

To aid in ADR detection, many national reporting tools and social support networks have been developed. These can be accessed online by patients, practitioners, and researchers alike. Self-reported patient information, in particular, captures a valuable perspective that might not be captured in any other way. The information voluntarily submitted by patients to national agencies, like the US FDA’s MedWatch program or the UK MHRA’s Yellow Card Scheme is estimated to reflect less than 10% of the adverse effect occurrences (Inman & Pearce, 1993; Yang et al., 2012). Thus, critical ADRs may go undetected until the harm done grows to a noticeable level.

Social media networks can present an alternative to formal national agency sites, and could prove to be a promising resource for ADR detection. Patients often submit pharmaceutical information in a wide variety of social networking resources, such as disease specific communities, blogs, microblogs, public news websites, or drug discussion forums. There has been a significant interest in these alternative sources for ADR detection in the last few years, starting with Leaman et al. (2010), which focused on comments from Daily Strength¹, a health-related community forum.

Twitter as a source of such comments, in general, presents different challenges. Here, we provide a manually annotated corpus from Twitter and outline these challenges and the methods we used to mine the Twitter microblogging platform for ADRs. The corpus contains 10,822 tweets annotated by domain experts for the presence of an ADR (as a binary attribute). In addition, it contains annotations for spans of text referring to specific ADRs along with UMLS IDs for the concepts.

Natural language processing from social media text is a very challenging problem. The text is often unstructured and informal, and may contain a large number of misspelled words. For health-related mining, the challenges compound. Specific to our problem, identifying social media postings that reference a prescription drug is just the beginning of the problem; colloquialisms, hypotetical postulation, or information not relevant to personal experiences need to be filtered. For example, users might re-tweet news reports or comments about side effects heard on a television commercial. Even if the data were structured and formal, automatic identification of ADRs is itself a challenging task. This is due to the complex relationships between drugs and their indications, adverse effects, and beneficial effects, as well as the great diversity of informal and creative terms that can match adverse-effect lexicon terms. In other words, if a patient simply mentions that a drug “makes them sleepy” it may apply to the drug’s treatment effectiveness (as a

¹http://www.dailystrength.org/

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number 1R01LM01176. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health.”
sleep aid), a beneficial effect (positive, but unintended effect), or an ADR (adverse effect). Furthermore, manually annotating data to structured vocabulary codes located in a dictionary (lexicon codes) increases complexity. Mentions like “sleepy”, “tired”, “groggy”, and “excessive sleepiness” all have different lexicon codes (described in section 3.3) without normalization, and consequently, obscure the data and diminish automatic learning model accuracies.

We focus this paper on a description of the corpus and the process followed to monitor Twitter for comments related to a drug. The correct acquisition of tweets (taking into account misspellings, for example), adequate pre-processing, and systematic manual annotation are paramount to optimize the performance of machine learning methods trained using this data. We present a benchmark text mining application (classification of tweets as to whether they include an ADR mention or not) to demonstrate the potential utility of the corpus. Classification was performed using two classes of machine learning algorithms: Naïve Bayes (NB) and Support Vector Machines (SVM).

2. Related Work

In the past, a handful of different social media resources have been used in detecting ADR information. We can broadly group the efforts into those trying to exploit the data produced by online communities and disease-specific blogs, and those attempting to mine social media sites such as Twitter. The differences between them have not been well-studied, but data from online communities is generally more structured, since comments are usually grouped by treatment and/or disease. For example, two articles focused solely on data from a diabetes forum are discussed in the following subsection. We outline related work for each of these two broad categories, to better appreciate the differences.

2.1 Online Communities and Blogs

A few studies have explored the potential of the postings on the Daily Strength online community to detect ADRs. The work by Leaman et al. (2010), presented a lexicon based approach to detect ADRs. In addition, they compared the frequencies of ADRs in user comments to the frequency of documented ADRs. They explained that the most common sources of errors arise from novel/creative phrases, idiomatic expressions, string matching problems associated with misspellings, ambiguity, and miss-categorizations (identifying an indication as an ADR).

In order to address some of the limitations of the lexicon-based approach, Nikfarjam & Gonzalez (2011) proposed a method to capture the underlying syntactic and semantic patterns from the Daily Strength reviews. This study benefited from a large, manually annotated corpus (1,200 records) and an algorithm that can detect expressions not included in a lexicon. However, there are some limitations associated with the pattern-based method that prevent it from being a stand-alone solution for this task. One of the main challenges is its dependence on the size of the training data, since it identifies an extraction pattern only if enough matching sentences are observed in the training data.

Two different studies were published in 2013 that utilized diabetes focused data. Akay, Dragomir & Erlandsson (2013) developed a methodology (text mining and self-organizing maps) to correlate positive and negative word cluster groups and with medical drugs and devices. This could aid the detection of ADRs through preprocessing for semantic tone/drug relationships prior to more complex ADR analysis.

Likewise, Liu & Chen (2013) utilized the diabetes forum from Daily Strength (different than the treatment comments used by the above), and created a platform called AZDrugMiner that can be used on a variety of online blogs. This framework was evaluated with manually annotated forum posts and broken down into four different methodologies: medical entity extraction (a lexicon based approach utilizing MetaMap), adverse drug event extraction (transductive SVM classifier on labeled an unlabeled data), report source classification (SVM classifier to differentiate personal experience from hearsay using labeled and unlabeled data), and analysis of ADR reports in the FDA Adverse Event Reporting System (FAERS) vs. reports in patient forums (similarities and differences for various drugs).

Yang et al. (2012) analysed the MedHelp health community for drug safety signal detection. They relied on lexicon based ADR matching, association mining, and calculation of the proportional reporting ratio. Their study was limited to 10 predetermined drugs and 5 predetermined ADRs. Thus, they did not identify unknown ADRs without prior conditions — the authors simply looked for presence of pre-specified ADRs.

2.2 Twitter

Twitter, one of the largest social media websites, has over 645,000,000 users (as of January 1, 2014) and grows by an estimated 135,000 users every day, generating about 9,100 tweets every second — a potential gold mine of information for researchers interested in studying population trends. This is especially true given Twitter’s application programming interface (API), which makes part of its data publicly available and easily accessible. A recent survey revealed that 26% of online adults discussed personal health issues and that 42% of them use social media to post or seek information about health conditions (Parker et al., 2013). Tempting as it is to assume this is all easy to access and process, utilizing this enormous amount of succinct, unstructured, and informal information in a

---

way that can be useful to public health officials goes beyond the inherent challenges of natural language processing of such data, but automatic processing and extraction is a necessary first step. We outline here the advancements in the extraction of adverse drug reactions from twitter in the past two years. Nakhasi et al. (2012) developed a manually annotated corpus of 770 tweets for patient safety event mentions that were preventable, adverse, care related, explicitly the result of health care actions/procedures, and something experienced on a first hand or someone personally known. The annotators noted who reported the event, the error source, error type, and error emotion. The results indicated that as much as 22.2% of the errors were medication related. However, this corpus was not intended to develop a natural language processing tool or algorithm to extract information — it was an annotated corpus and analysis of its promising statistical value. Bian et al. (2012) created an SVM classification model from a large dataset — 2 billion tweets — on a high performance computing platform. The model sought to find prescription drug users and potential adverse events using 5 investigational cancer drugs. All data from a single user was aggregated into one document for temporal analysis by the SVMs — this ensured analysis of comments that spanned multiple tweets. Overall, 239 users were annotated for personal relationship to drug effects and 72 (the positive cases for personal relationships) were annotated for adverse events (resulting in 27 adverse events). Since the performance was limited, the authors suggested a number of error sources: noise (fragmented sentences, misspellings, non-word terms, odd-abbreviations, etc.), errors in tagging by MetaMap (Aronson & Lang, 2010), errors due to non-standard terms, and part-of-speech tagging errors due to highly unstructured text. In another study (Jiang & Zheng, 2013), 5 drugs with established market presence were studied, collecting 6,829 tweets, and a classification model for drug effects was developed. Three classification models were analysed, with personal pronouns and sentiment analysis, to determine if tweets regarding "personal experiences." MetaMap was used to evaluate drug effects, which included diseases, findings, injuries, dysfunctions, and symptoms. The findings were compared to information on PatientsLikeMe6 and MedLinePlus7 with a 74-86% matching rate. However, the corpus studied was not annotated by domain experts and the study was not focused on ADRs. Our corpus has broader coverage than the ones noted, with 74 carefully selected drugs queried, including misspellings, with a total of 10,822 tweets manually annotated by experts. It includes both binary annotation for classification applications, and specific concept annotation with mappings to UMLS concept IDs for the approximately 1,200 tweets that include a mention of an adverse reaction, indication, or beneficial effect. This can facilitate the development of advanced concept extraction and identification techniques for ADRs in Twitter.

3. Methods

Our manually annotated corpus was created through extraction of tweets related to 74 drugs of interest, using their brand and generic names, and phonetic misspellings. These were annotated for the presence of ADRs (a binary attribute for each tweet), the location/span of the reaction mentions, and the UMLS concept IDs for the ADR mentions. The overall process flow is shown in Figure 1 and subsequently described.

Figure 1: Data collection and annotation flowchart.

3.1 Drug List Generation

As part of a larger study, we selected a set of drugs to be monitored for different kinds of adverse effects. It includes both generally used drugs whose adverse effects are well known (truth set), and drugs released between 2007 and 2010 for which not all adverse effects are yet known and will only become visible as the drugs are more widely used. Drugs in the truth set were selected on the basis of their widespread use, as demonstrated by their presence in the Top 200 products by volume in the U.S. market. Many of the drugs for the truth set have come into widespread use recently, which allows for testing the capability of the natural language processing so at the time of release. For the newer drugs, going back to 2007 allows for market growth leading to common prescribing and comments on the site. The list was narrowed based upon forecasts for widespread use, the prevalence of disease states and conditions, and on whether the drug was new in class. Major categories

---

6 http://www.patientslikeme.com/
7 http://www.nlm.nih.gov/medlineplus/
include drugs for the central nervous system and neurological conditions such as Alzheimer’s disease and schizophrenia. Treatments for age-related diseases like diabetes, cardiovascular diseases, urinary dysfunction, and musculoskeletal disorders also met the criteria for potential widespread use, given our life expectancy.

The list of drugs used as keywords to monitor Twitter was first expanded by including the generic and brand names of the drugs. Then, we extended the drug list to include misspelled drug names. This was critical to obtaining relevant tweets, as drug names are often misspelled in social media. We generated the misspellings through a ‘phonetic spelling filter’ (Pimalkhute et al., 2013). This gave preference to variants that reflect the phonemes of the correct spelling. Examples of variants and misspelled tweets are shown in Table 1. Initially, the tool generated a large number of misspellings, out of which 18% were added to the list of drug names. This percentage was experimentally determined to maximize the tweet coverage while minimizing the number of terms needed to query Twitter. This is important because twitter API allows only 400 keywords per application key. This technique allowed us to capture an estimated 50 to 56% of tweets mentioning the drug.

<table>
<thead>
<tr>
<th>Original Drug Name</th>
<th>Example variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prozac</td>
<td>prozaac, prozax, prozaxe,</td>
</tr>
<tr>
<td>Paxil</td>
<td>paxl, pxil, paxol.</td>
</tr>
<tr>
<td>Seroquel</td>
<td>seroquels, seroqul, seroqual</td>
</tr>
<tr>
<td>Olanzapine</td>
<td>olanzapin, olanzapoine, olanzoaione</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example Tweets with Seroquel Spelling Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>@Psychological HA! Not if you’re on # Seroquilt</td>
</tr>
<tr>
<td>@BipolarBlogger did you ever try the Seroquilt XR???</td>
</tr>
<tr>
<td>Gone from 50mg to 150mg of Seroquilt last night. Could barely wake up this morning and I feel like my body is made of lead</td>
</tr>
<tr>
<td>@AndrewH_Smith Is the Inderal helpful? And yeah, they are short lasting but non addictive. You could try Seroquelt too but it’s pretty strong</td>
</tr>
</tbody>
</table>

Table 1: Examples of spelling variants that were generated and tweets that contained drug misspellings.

3.2 Preprocessing

After finalizing the drug name list, we accessed the twitter database through their publicly available API. The API allowed us to obtain matching tweets up to a volume equal to the streaming cap (~1% of all public tweets), restricted to 1000 requests per day. This resulted in 187,450 tweets over a period of 6 months, which were then filtered to remove advertisements. To remove advertisements, we removed tweets that contained URLs. This cut down the tweets to 71,571.

Next, we balanced the dataset to select a set for annotation. This helped prevent dominance of some drugs over other drugs, as some drugs are much more popular than others. For example, 58,000 tweets were about nicotine; while other drugs averaged 240 tweets. We randomly selected a maximum of 300-500 tweets per drug, for a total of 10,822 tweets. All twitter datasets were stored and retrieved for later use using a Mongo9 database for access by the annotators.

3.3 Manual Annotation

We sought to annotate our corpus not only for the presence or absence of ADR mentions, but also to identify the span of the expressions conveying individual ADRs, and to map them to formal medical terminology (i.e., assigning them UMLS concept IDs). For the purpose of annotation, an ADR was defined as: “an effect of the drug, which is not desired, and includes mentions that described a worsening of the patient’s initial health condition”. It was important to distinguish these from indications, defined as: “the causal condition, symptom, or disease that was the reason for the patient taking the drug”. Both these types of concepts are specifically annotated in the corpus. The distinction between ADRs and indications highlights the benefits of manual annotation as opposed to dictionary matching since the two types share concept names and UMLS codes, and would be indistinguishable if not by context.

The lexicon used to select the UMLS IDs was developed by augmenting a prior lexicon used in Leaman et al., (2010). Originally, the lexicon included groups of terms from COSTART10, SIDER Version I (Kuhn et al., 2010), MedEffect11, and a limited list of idiomatic expressions. The augmented lexicon used here also includes terms from SIDER Version II (Kuhn et al., 2010) and the Consumer Health Vocabulary (CHV) (Zeng-Treitleit et al., 2008). We narrowed the scope of the terms in the UMLS to signs or symptoms, and excluded terms relating to topics like medical procedures. Overall, this resulted in 7,483 unique terms (UMLS concept IDs) and 16,182 concept names, which are normalized to the unique terms.

Next, two annotators manually annotated the processed tweets. The corpus comprises two types of data: binary annotation of ADRs and full ADR annotations (specified span and UMLS concept IDs). Chronologically, the annotators first analysed all 10,822 the tweets for the binary annotations. Following that, the tweets with ADRs were separated for full annotation. Both binary and full annotations were performed using an unpublished tool developed in-house.

9 https://www.mongodb.org/

9 10 11
Annotators held weekly meetings to discuss the annotated tweets, correctness of concept labels, and develop annotation guidelines. The two annotators have medical or biological science background. Some meetings also included the full project team (one biomedical informatics student with a computer science background, two computer science students, and a pharmacology doctor). They annotated a total of 10,822 tweets, utilizing the following general principles for the concept annotation:

- Location boundaries of every mention should be minimized but the boundaries must also capture the entire concept.
- Every annotation should be normalized to a UMLS concept ID that most closely matches the meaning.
- For indirect matches, the most general ID should be used (may require annotator deliberation).

For instance, “weight gain,” “gained 20 pounds”, “put on too much weight”, or “fat fat fat” would all be annotated to a general concept ID for “weight gain”. Instances that caused confusion in selecting the most general term were caused confusion in selecting the most general term were resolved by a general concept ID for "weight gain". Instances that caused confusion in selecting the most general term were discussed during meetings. For more information on the annotation process, please refer to the annotation guideline that accompanies the corpus.

3.4 Binary Classification

To demonstrate the utility of the corpus in detecting the tweets containing ADR mentions, we used NB and SVM classifiers for the binary classification task. Our intent was to investigate if supervised machine learning models can be trained on the annotated data in our corpus so that the tweets containing ADR mentions can be automatically identified. We used two classes for the experiments: hasADR and noADR — representing instances that contain ADR mentions and those that don’t, respectively.

We performed preprocessing of the text to reduce noise. All of the words were lemmatized and transformed to lower case letters. Forms were also normalized. We implemented normalization through a dictionary algorithm that expands abbreviations (from a manually created list of commonly used abbreviations). For instance, “abt” was normalized to “about”. We also implemented an algorithm similar to the one proposed by Brody & Diakopoulos (2011) for reducing word length in regards to emphasis words. For example, words like “coolllll” or “cooooooool” were reduced to “cool”.

For modeling text as vectors, we represented text instances as vectors in space of vocabulary (defined as set of all the unique words in the corpus). We used a simple term frequency scheme for vectorization, where value of it feature in the vector is equal to the number of times that feature or word occurs in that particular instance. For the SVM classifier, we used a linear kernel, and made no attempts at parameter optimization.

The binary dataset is very imbalanced and the number of noADR instances is much greater than the hasADR instances; thus, we trained the classifier on a more balanced dataset using randomly selected noADR instances. We created three sub-datasets varying in the skewedness towards the noADR class. The first dataset (dataset-1) is a balanced dataset, implying equal distribution of both classes: 1,008 hasADR, 1,008 noADR. The second dataset (dataset-2) has 60% noADR instances (1,008 hasADR, 1,512 noADR). The third dataset (dataset-3) has 70% noADR instances (1,008 hasADR, 2,352 noADR).

4. Results and Discussion

After developing the corpus, we analyzed our results to better understand its attributes and potential usefulness for text mining applications. We looked at the frequency and distribution of ADR mentions, the agreement between annotators (Table 2), and the performance on text mining classifiers.

4.1 Corpus Description and Statistics

A total of 10,822 tweets were annotated for the presence of ADRs by two experienced annotators, yielding approximately 1,200 tweets with at least one ADR mention (one annotator reported 1,008, while the other reported 1,255 tweets with 1,256 and 1,436 ADRs mentioned, respectively). For the experiments described in this paper, we considered the annotations by the first annotator to be the gold standard. In the final version of the corpus, the disagreements in the annotations will be resolved by Dr. Karen Smith.

To quantify the disagreements between annotators, we compared their annotations using partial matching criteria for span, and exact matching for concept IDs and the binary annotation. We report precision, recall and F-measure (Table 2) for agreement. For concept spans, an agreement (TP) occurs if there is some overlap on what the two annotators mark as the span for a concept. For concept IDs, annotators had to map an annotated concept to semantically equivalent UMLS concept IDs for an agreement to occur. For binary annotations, we computed the inter-annotator agreement (IAA) using Cohen’s kappa (Carletta, 2006) and obtained a value of 0.69. According to Viera & Garrett (2005), a kappa of 0.61-0.80 indicates “substantial agreement”.

As expected, concept IDs presented a larger source of disagreement than the spans of the ADR mentions within the text. However, we obtained fairly high values overall, especially for recall.

<table>
<thead>
<tr>
<th>Comparison Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F–Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Span</td>
<td>0.810</td>
<td>0.980</td>
<td>0.887</td>
</tr>
<tr>
<td>Concept ID</td>
<td>0.728</td>
<td>0.969</td>
<td>0.831</td>
</tr>
<tr>
<td>Binary</td>
<td>0.816</td>
<td>0.883</td>
<td>0.845</td>
</tr>
</tbody>
</table>

Table 2: IAA for concept span, concept ID, and binary (presence or absence of ADRs) annotations.
4.2 Binary Classification Results

To evaluate the performance of the classifiers, we computed evaluation metrics (precision, recall, f-measure and accuracy) using 10-fold cross validation. Table 3 shows the performance of the NB and SVM classifiers for binary classification. The equations for the metrics are shown in Equations 1-4.

<table>
<thead>
<tr>
<th></th>
<th>noADR</th>
<th>hasADR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Dataset – 1</td>
<td>0.608</td>
<td>0.840</td>
</tr>
<tr>
<td></td>
<td>0.799</td>
<td>0.684</td>
</tr>
<tr>
<td>Dataset – 2</td>
<td>0.670</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>0.845</td>
<td>0.739</td>
</tr>
<tr>
<td>Dataset – 3</td>
<td>0.745</td>
<td>0.883</td>
</tr>
<tr>
<td></td>
<td>0.894</td>
<td>0.796</td>
</tr>
</tbody>
</table>

Table 3: Classification results for the NB and SVM classifiers on the three data sets.

\[
\text{precision} = \frac{TP}{(TP + FP)}
\]

\[
\text{recall} = \frac{TP}{(TP + FN)}
\]

\[
F\text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})}
\]

\[
\text{Accuracy} = \frac{\text{Total correct predictions}}{\text{Number of test instances}}
\]

Equations 1-4: Formulaic equations for the classifier evaluation metrics.

4.3 Error Analysis

4.3.1 Corpus and Annotation

The disagreements in annotations between the annotators were analysed for the three categories: concept ID disagreement, span differences and nonmatching annotations. The largest source of disagreement came from the differences in concept IDs selected. There are several key reasons behind these discrepancies. The first is the similarity, and in some cases exact matches, in terms found in the lexicon associated with different concept ID numbers. For example, a patient may mention a “hangover” as an ADR. In the lexicon there is a concept ID for “pill hangover” and another ID for “hangover from alcohol or any other drug substance”, the similarities of the terms make it difficult to ensure agreement as there is no obvious reason to select one over the other. Another source of discrepancies in concept ID annotations comes from differences in the annotators’ interpretation of the idioms, slang or euphemisms used by the patient. The relatively small size of each tweet exacerbates this problem because it can eliminate the clues to meaning that can often be found in the context of large text documents.

The span and nonmatching annotation discrepancies mostly arise from the same issue, i.e., differing interpretations between annotators regarding how much of the text should annotated to capture the concept. This can be problematic at two levels: not only does it lower IAA, but the span selected by an annotator tends to influence the concept ID selected, further adding to disagreements. For example, annotating “tremors in hands” vs. just “tremors” resulted not only in span disagreement but also in concept disagreement with the first being mapped to the ID for “tremor of hands” and the latter to “tremors, shaking”. Nonmatching annotations are instances where one annotator annotates a portion of the text that was not annotated at all by the other annotator. Disagreements of this nature often highlight the difficulties annotators can have in determining whether the person is discussing their own adverse experience with the drug or if they are merely providing a commentary on the drug. One such case is the following tweet: ‘depression hurts, cymbalta can help? one of that s*** many awful symptoms is thoughts of suicide...’. One annotator selected “thoughts of suicide” as an ADR and the other did not, based on the criterion that an ADR was to be annotated only if it was
experienced by the user posting the mention. The first annotator interpreted the message as such, while the second did not. Once again, the limited nature of tweets removes contextual clues and annotations become a subjective decision made by the annotators. Future work to improve IAA might include the review and revision of the lexicon to unify concepts that are the same or similar. To improve issues with span selection, ongoing annotation meetings to discuss and revise the annotation guidelines with the purpose of improving clarity and conciseness of instruction should help annotator decisions in the future.

4.3.2 Binary Classification
To obtain estimates about the relative performances of the NB and SVM classifiers on this data, we compared their accuracies against the majority labeling baseline. Majority labeling is the method of labeling every test instance to the majority class in the test set. This evaluation enabled us to (i) compare the performances of the classifiers against a simple baseline that does not use lexical data from the corpus, and (ii) compare the performances of the classifiers for balanced vs. unbalanced data sets. Figure 2 illustrates the results.

From the Figure, it can be observed that both classifiers perform better than the majority labeling baseline, even when the data set is heavily imbalanced (i.e., the majority class represents 70% of the data). This clearly indicates that the data and annotations in our corpus aid the training of automatic models for classification of ADRs. It can also be observed that the classifier accuracies increase slightly as the proportion of the majority class increases. Our inspection of the results show that this is because as the number of instances for the majority class increases, so does the accuracy over the majority class. Since the overall accuracy relies more on the majority class as its proportion increases, increase in the classification accuracy for the majority class results in overall accuracy increase as well. This, however, does not mean that the accuracy for the hasADR class increases as well as the imbalance in the data increases. The overall effect of the training set proportions on the classification accuracy for the hasADR class requires further experimentation. In this analysis, we did not attempt to deeply analyze how the training and test set proportions affect classification accuracy, or what training-test ratio would be ideal for the classification of real-life data. We leave this as future work.

We analyzed the false positives associated with the hasADR. In many of the instances, the author would talk positively about a drug, but our supervised learning algorithms, based on the presence of specific terms, classify the statements as ADR mentions. For example: “my [drug] is kicking in, I can feel it.” Such errors may perhaps be eliminated by using more intelligent feature selection techniques for the classifiers. A number of false positives contain mentions from the ADR lexicon, but the statements do not actually report adverse effects.

Instead they present the authors’ questions (e.g., “who’ve been prescribed [drug] for sleep? Has it helped at all?”), or just life stories (e.g., “I could pass out with a moments notice even with this [drug] in my system”). Another group of errors are tweets with idiomatic expressions and sarcasms, such as “I took a [drug] and yet I’m in an awesome mood. This never happens”. Many of the misclassifications are due to the lack of deep semantic analysis of the lexical contents of the tweets. However, such techniques are beyond the scope of this paper as our intent is to present the corpus and investigate its potential for the implementation of advanced automatic techniques for ADR detection. The performances of our classifiers using basic lexical features are very promising. We leave the implementation of more complex techniques for classification as future work.

5. Conclusions
We presented an annotated Twitter corpus focused on ADR mentions with broad pharmacological coverage, collection twitter comments about 76 drugs. Only 65 of the drugs had one or more associated tweets, with wide variability in the number of tweets per drug. We selected a balanced number of tweets per drug, to form a corpus that contains a total of 10,822 tweets manually annotated by experts. It includes both binary annotation for classification applications, and specific concept annotation with mappings to UMLS concept IDs for the
approximately 1,200 tweets that included a mention of an adverse reaction, indication, or beneficial effect. This can facilitate the development of advanced concept extraction and identification techniques for adverse drug reactions in Twitter. The binary annotations are available for download at http://diego.asu.edu/downloads. We show the utility of our corpus by applying two supervised machine learning approaches for the binary classification task of identifying if tweets contain ADR mentions or not. Although the classifier performances are modest, it is intended as a baseline for future development. Multi-stage natural language processing platforms could be applied for the binary classification and other associated ADR detection tasks. We applied the NB and SVM classifiers using surface level lexical features (i.e., word vectors). Our goal is to incorporate features through the use of deep semantic analysis of the text associated with the tweets.

References


