# Quantifying Self-Reported Adverse Drug Events on Twitter: Signal and Topic Analysis

Vassilis Plachouras
Thomson Reuters
Research & Development
1 Mark Square
London, EC2A 4EG, UK
vassilis.plachouras@tr.com

Jochen L. Leidner
Thomson Reuters
Research & Development
1 Mark Square
London, EC2A 4EG, UK
jochen.leidner@tr.com

Andrew G. Garrow
Thomson Reuters
IP & Science
77 Hatton Garden
London, EC1N 8JS, UK
andrew.garrow@tr.com

## ABSTRACT

When a drug that is sold exhibits side effects, a well functioning ecosystem of pharmaceutical drug suppliers includes responsive regulators and pharmaceutical companies. Existing systems for monitoring adverse drug events, such as the Federal Adverse Events Reporting System (FAERS) in the US, have shown limited effectiveness due to the lack of incentives for healthcare professionals and patients. While social media present opportunities to mine information about adverse events in near real-time, there are still important questions to be answered in order to understand their impact on pharmacovigilance. First, it is not known how many relevant social media posts occur per day on platforms like Twitter, i.e., whether there is "enough signal" for a post-market pharmacovigilance program based on Twitter mining. Second, it is not known what other topics are discussed by users in posts mentioning pharmaceutical drugs.

In this paper, we outline how social media can be used as a human sensor for drug use monitoring. We introduce a large-scale, near real-time system for computational pharmacovigilance, and use our system to estimate the order of magnitude of the volume of daily self-reported pharmaceutical drug side effect tweets. The processing pipeline comprises a set of cascaded filters, followed by a supervised machine learning classifier. The cascaded filters quickly reduce the volume to a manageable sub-stream, from which a Support Vector Machine (SVM) based classifier identifies adverse events based on a rich set of features taking into account surface-textual properties, as well as domain knowledge about drugs, side effects and the Twitter medium. Using a dataset of 10,000 manually annotated tweets, a SVM classifier achieves F1=60.4% and AUC=0.894. The yield of the classifier for a drug universe comprising 2,600 keywords is 721 tweets per day. We also investigate what other topics are discussed in the posts mentioning pharmaceutical drugs. We conclude by suggesting an ecosystem where regulators and pharmaceutical companies utilize social media to obtain feedback about consequences of pharmaceutical drug use.



*SMSociety '16, July 11-13, 2016, London, United Kingdom*



DOI: http://dx.doi.org/10.1145/2930971.2930977

## CCS Concepts

•**Networks** → **Social media networks;** •**Information systems** → *Clustering and classification;* Data analytics;

## Keywords

Drug side effects, Classification, Text Mining, Natural Language Processing, Twitter

## 1. INTRODUCTION

According to the World Health Organization (WHO), "pharmacovigilance is the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems" [29]. Identifying Adverse Drug Reactions (ADRs) as early as possible is very important in order to improve the well-being of people and to reduce the additional costs for health systems to treat patients from ADRs.

Clinical trials are required to identify ADRs in the pre-marketing phase of drug development. The limited number of participants in clinical trials and constrained time period, however, do not guarantee that all the ADRs of a new drug will be identified. After the approval of a drug by regulators, monitoring for adverse drug events (ADEs) continues through the submission of reports by pharmaceutical companies. Healthcare professionals can also submit reports about adverse events and some countries operate systems that allow patients to directly report ADEs (path indicated by black arrows in Figure 1). The investigation of the submitted ADE reports may lead to the identification of previously unknown ADRs if a causal relationship is established between the use of a drug and the response to it. Existing schemes are limited by the substantial under-reporting of adverse events [10], due to the lack of incentives for healthcare professionals and patients to submit reports.

Recently there has been a growing body of research work aiming to mine user-generated content and social media to identify adverse drug events [12]. Mining social media enables both regulators and pharmaceutical companies to listen to what patients say about drugs in real time, as indicated by the grey arrows in Figure 1. However, there are still several open questions to assess the impact of using social media for pharmacovigilance. First, there is no indication of the number of posts that mention adverse drug events. Second, while detecting adverse events is the main objective for pharmacovigilance, there are other related topics, such as drug efficacy, the presence of which has not been quantified.

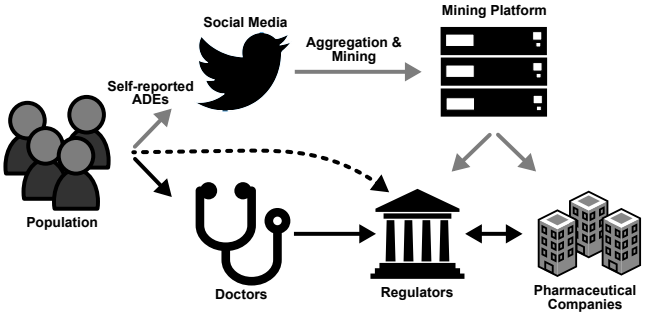


**Figure 1: The ecosystem of pharmaceutical companies, regulators, doctors, patients and social media.**

In this work, we present the results from a study of a large-scale system that mines Twitter[1] to identify adverse events. While Twitter provides an opportunity to identify adverse events in near real-time from the tweets posted by users, it is a challenging medium due to the volume of data and the characteristics of the text in tweets. Only a small fraction of the 500 million tweets posted daily are related to health issues [22]. If tweets about adverse events happen only rarely (e.g., once per year), the attempt to mine them would not make economic sense. However, as we will show, there is a regular occurrence of adverse event mentions. The system we introduce collects tweets through a set of cascading filters which aim to keep the tweets that are most likely to mention adverse events. Then, the system applies a Support Vector Machine (SVM) classifier to identify tweets that mention adverse events. The classifier has been trained and evaluated in a supervised setting using 10,000 manually annotated tweets.

In this paper we address the following research questions:

**RQ1**: How effectively can tweets mentioning adverse drug events be identified automatically?

**RQ2**: What is the signal yield of the developed system, i.e. how many tweets are labeled as positive by the classifier?

**RQ3**: What other topics are discussed in tweets mentioning pharmaceutical drugs?

The main contributions of this work are the following:

- **Automatic system:** We introduce a system for mining Twitter to find adverse effects of drugs. Our system uses a number of cascading filters combined with a machine learning classifier and a number of novel surface-based and linguistic features to identify the most relevant data for further processing.
- **Yield volume analysis:** We have established that compared to the 500 million tweets posted daily, the output of our system is in the order of hundreds of tweets per day, which can be analyzed and assessed by a single human analyst.
- **Topics discussed in tweets:** We identify other topics discussed by Twitter users in posts mentioning pharmaceutical drugs.

[1]http://www.twitter.com/

The remainder of this paper is structured as follows. In Section 2, we characterize the data collection and describe the manual annotation process. In Section 3 we present the architecture of the developed system, and in Section 4 we describe the features of the adverse event classifier. In Section 5 we report the evaluation results. Sections 6 and 7 present the analysis of the signal yield of the classifier and other topics discussed in relevant tweets, respectively. Finally, Section 8 reviews related work and Section 9 summarizes and concludes with some suggestions for future work.

## 2. DATA

There are more than 500 million tweets posted by users every day. Only a small fraction of these tweets mention the names of pharmaceutical drugs, their active substances and potential side effects of these drugs. In order to train a classifier to identify tweets that mention adverse events, we have collected a set of tweets in which at least one pharmaceutical drug is mentioned. The data collection was performed using the Twitter Search API[2], by repeatedly submitting queries with a delay such that we respect the rate limit of the API. We used the following keyword lists:

- A list of 397 blockbuster pharmaceutical drugs, corresponding to the names of the best selling drugs in terms of revenue in excess of US $200 million in 2012, according to Thomson Reuters Cortellis[3].
- A list of 115 hypertension drug names.
- A list of six often-mentioned drugs (abilify, accutane, cipro, seroquel, wellbutrin, zopiclone) and a list of phrases indicating side effects (anxiety attack, appetite, bleed, bone pain, constipation, cotton mouth, dizzy, drooling, drowsy, dry mouth, faint, fatigue, gain weight, hallucination, heart disease, hives, hypertension, itchy, joint pain, malaise, memory loss, mood swing, nausea, nightmare, palpitation, panic attack, put weight, vomit, weakness), selected after collecting tweets using a list of more than 1,300 pharmaceutical drugs. Note that accutane is the name of a drug which was used until 2009 to treat severe acne. It has been withdrawn since then, but its generic versions are still commonly referred to with the same name.

In this work we did not consider synonyms of drug names, or groups of drugs based on the active ingredients. We have used the list of blockbuster drugs in order to ensure that we collect a large number of tweets. Similarly, we have selected the six drugs and the phrases indicating side effects by observing that there was a large number of tweets matching them and mentioning adverse events. Finally, we have included the list of hypertension drug names in order to obtain tweets about a specific medical condition.

For the tweets retrieved with the blockbuster and hypertension drug lists, we have performed random sampling and selected the tweets to manually annotate. Regarding the list of six drugs and the list of side effect phrases, we have annotated all the matching tweets.

For the annotation of ground truth data, we categorized as positive examples all tweets that discuss an adverse event

[2]https://dev.twitter.com/rest/public/search
[3]http://lifesciences.thomsonreuters.com/products/cortellis

experienced by the user who posted the tweet. All other tweets were treated as negative examples for the purpose of training a binary classifier. In total, we have annotated 10,000 tweets with 1,432 positive and 8,568 negative examples. The human annotators went beyond just checking the presence of a drug and a side effect; they were tasked to focus on identifying only those tweets that mention drug side effect instances, not e.g. jokes or abusive language about drugs where the mention of side-effects are coincidental. We have measured the inter-rater agreement on a sample of 404 tweets annotated. The human inter-rater agreement level of $\kappa = 0.535$ is commensurate with the difficulty of the task, which arises from the use of colloquial expressions and the lack of context[4]. To resolve disagreements on the sample of 404 tweets rated by both annotators, one of the annotators made a decision after reviewing the corresponding tweets.

In addition to the ground truth data, we have also collected a large set of tweets based on a list of more than 2,600 pharmaceutical drugs. The collection of data has taken place from July 9 to September 4, 2014. In total we have collected 1.5 million tweets, which we use to estimate the *yield* (or signal-to-noise ratio) of our system in Section 6.

# 3. ARCHITECTURE

In this section, we first outline the architecture of the system (Figure 2) and provide a description of its modules. The details of the linguistic extraction and classification features are provided in Section 4.

## 3.1 Topic and Volume Filters

The stream of tweets is processed by the Topic Filter, which keeps only tweets having at least one keyword from the set of keywords related to drugs (described in Section 2). Next, we apply the Volume Filter, which allows a tweet to pass if all the following conditions are satisfied:

1. It is not a re-tweet without modifications (forwards without edits), because such re-tweets do not add any primary information about adverse events. Re-tweets with edits are not discarded by the Volume Filter.

2. It does not contain a hyperlink to a Web page. This is motivated by the observation that link-containing posts are typically commercial offerings. We are interested in personal experiences self-reported by patients, unlikely to contain hyperlinks.

3. It is written in English. Our method can be applied to other languages in principle, but that is beyond the scope of this work.

The Volume Filter applies these checks in order of decreasing throughput. For example, we first check whether a tweet is actually a re-tweet. Last, we perform language detection to identify whether the tweet is written in English [17][5]. In an operational setting, the Topic Filter and Volume Filter can be executed on Twitter's firehose.

[4]Unlike [14], we refrain from referring to any level of agreement as "good" or "fair" in an arbitrary way, especially given the criticisms of [27] and [8].

[5]https://github.com/carrotsearch/langid-java

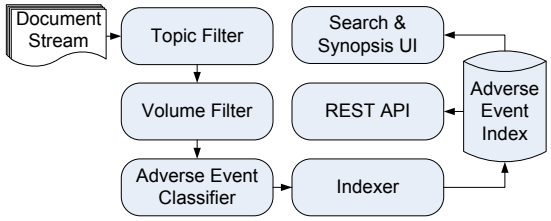


**Figure 2: The architecture of our system.**

## 3.2 Adverse Event Classifier

Tweets that pass the Volume Filter are processed by a binary classifier which is trained to distinguish whether a tweet mentions an adverse event or not. We learn to classify tweets discussing adverse events in a supervised learning setting with a Support Vector Machine (SVM). In our setting, a pattern $x \in \mathbb{R}^d$ is a $d$-dimensional vector of features extracted for a tweet. We associate each pattern $x$ with a response value $y \in \{-1, +1\}$, where $+1$ indicates that the tweet corresponding to the pattern discusses adverse events and $-1$ indicates otherwise. Before training and evaluating the SVM classifier, we normalize the feature values such that they are in the range $[0, 1]$. We use the LIBSVM implementation of SVM [3] with a linear kernel.

## 3.3 Indexer

The Indexer reads the tweets that have been classified as positive and adds them to the Adverse Event Index. For each tweet, the index maintains fields for the drugs found in the tweet, the adverse events, the timestamp of the tweet and its text.

## 3.4 REST API

The index of tweets is accessible via an API which provides two endpoints to query the data. The first one enables the retrieval of tweets based on the mentioned drugs, side effects, or any word in the text of tweets. Matching tweets can be further restricted within a given time range, or by filtering tweets where no side-effect from our gazetteers matched. The second endpoint returns information about the occurrences of drugs and side effects in the data, or other facets related to the gazetteers we have employed in the system.

## 3.5 Search and Synopsis User Interface

We have developed a Web-based application that enables users to search for drugs and adverse events, and explore the relations between them. The search interface provides information about the frequency of drugs and side effects in the search results, the volume of tweets over time, and it also displays a redacted version of the tweet text. A user can provide feedback about whether a tweet has been classified correctly or not. Figure 3 shows a screenshot of the Search UI. The Synospis UI provides a visualization of all the identified relations between drugs and adverse events.

# 4. FEATURES

We have developed a range of features based on the textual content of tweets as well as domain-specific resources. In the remainder of this section we describe the five groups of features we have extracted.

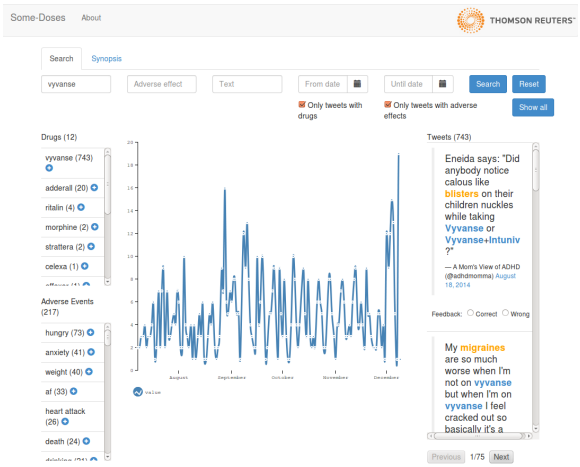


**Figure 3: Screenshot of the search GUI and results for the drug Vyvanse.**

## 4.1 N-grams

We extract all unigrams and bigrams from the text and keep the ones that contain only alpha-numeric characters. The text of tweets is first tokenized and normalized by lowercasing. We do not remove stopwords or stem tokens. The features we employ are the following:

- BIN_NGRAM$L$_$w$: a binary indicator feature, which is set equal to 1 if the tweet contains n-gram $w$ with length $L$, otherwise 0.

For example, for the text "I took two pills" for $L \in \{1, 2\}$ we generate the set of unigrams $\{\text{i}, \text{took}, \text{two}, \text{pills}\}$ and the set of bigrams $\{\text{i\_took}, \text{took\_two}, \text{two\_pills}\}$.

Our preliminary experiments have shown that using trigrams or higher order $n$-grams did not improve the performance of the classifier. Chee, Berlin and Schatz [4] and Sarker and Gonzalez [24] employ up to trigrams to develop a classifier to extract adverse events from health-related forums and Twitter, respectively, but do not evaluate their effectiveness separately. Nikfarjam and Gonzalez [20] mine association rules between $n$-grams of POS tags and adverse events. In the context of sentiment classification, Wang and Manning [28] report that trigrams did not help. The effectiveness of $n$-grams as features has not been explored in the related work in the area of detecting adverse events from Twitter [1][6].

We have also experimented with term-frequency weighted $n$-gram features, which did not result in classification performance improvements. This is expected because tweets are short and words are less likely to repeat. We have not employed tf-idf weighted $n$-gram features, because the distribution of idf values is likely to be dominated by high values, due to the large number of infrequent words present in tweets [31].

## 4.2 Surface features

We also exploit text surface features to characterize a tweet. Simple text surface features can prove useful in extracting elements from the context of users: their emotional state, engagement in discussions with other users, or their attitude towards the health issue they experience. We introduce the following text surface features:

```
I   tried using prayer to cure my   headache , but tylenol worked much better .
PRP VBD   VBG   NN     TO VB   PRP$ NN       , CC  NN       VBD    RB   JJR    .
```

**Figure 4: The text of an example tweet and the POS tags assigned to each of the tokens.**

- The number of characters in the tweet divided by the maximum length in characters of a tweet. The longer tweets are more likely to be informative, while there are very short tweets with just few characters.
- The number of mentions (i.e., @Username) found in the tweet. The presence of user mentions in a tweet indicates that there is a conversation between users. Hence, even though we process each tweet independently, we are able to leverage evidence from the interaction between Twitter users.
- The maximum number of times a character is repeated inside a token. This feature will have a high value when a user emphasizes a word by repeating several times a character, for example writing "sleeeepy" instead of "sleepy."
- A binary feature which is set equal to 1 if the tweet contains at least one all-numerical token, such as in the phrase "I took 2 Benadryl's tonight."
- A binary feature which is set equal to 1 if the tweet contains at least one all-upper-case token, for example "HELP."
- A binary feature which is set equal to 1 if the tweet contains at least one title case (initial upper-case) token, for example the word "Accutane."
- A binary feature which is set equal to 1 if the tweet contains at least one token with mixed capitalization like "InterCity."

## 4.3 Part-of-speech (POS) tags

We develop a set of novel features based on POS tags assigned to tokens in order to encode information related to the grammatical structure, for example whether the user asks a question or makes a comparison. We employ the OpenNLP Maximum Entropy POS tagger[6] to add POS tags for each token and we add the following features:

- Past-Present verbs: A binary feature that indicates whether a tweet contains verbs both in past and present tense. The feature value is set equal to 1 if the tweet contains verbs $w_1$ and $w_2$ where $w_1$!=$w_2$ and $\text{POS}(w_1) \in \{\text{VB}, \text{VBD}, \text{VBN}\}$ and $\text{POS}(w_2) \in \{\text{VB}, \text{VBP}, \text{VBG}\}$, otherwise 0. This feature is intented to capture patterns of past actions leading to present consequences that are indicative of potential causality.
- Question tags: A binary feature, which is set equal to 1 if the tweet contains word $w$ for which $\text{POS}(w) \in \{\text{WDT}, \text{WP}, \text{WP\$}, \text{WRB}\}$, otherwise 0.
- Comparative-Superlative tags: A binary feature, which is set equal to 1 if the tweet contains word $w$ for which $\text{POS}(w) \in \{\text{JJR}, \text{JJS}\}$, otherwise 0.

[6]https://opennlp.apache.org/

- Verb signature: Concatenation of all verb POS tags in alphabetical order.

If we compute the POS tag features for the example tweet in Figure 4, the verb signature feature VB_VBD_VBD_VBG is set equal to 1. Moreover, the binary feature indicating the existence of comparative or superlative POS tags is set equal to 1 because the POS tag for the token "better" is JJR.

## 4.4 Gazetteers

We incorporate domain-specific knowledge with a set of gazetteers (lexicons) we have built. The gazetteers are grouped in three themes, namely user vocabulary, company and medical gazetteers.

The user vocabulary gazetteers comprise lists of words and phrases indicating abuse, humor, fiction, intake, efficacy, as well as patient feedback about a drug. The company gazetteers include lists of words related to commercial spam, commercial advertisements from pharmaceutical companies, financial and share price information, company news, and company designators.

For the medical vocabulary, we have constructed gazetteers related to human body parts, adverse effect synonyms, side effect symptoms, adverse events, causality indicators, clinical trials, medical professional roles, side effect triggers and drugs. Because patients' language in tweets is rarely the language of medical professionals (e.g., "heart racing" vs. "tachycardia"), we have also employed a list of 32,167 expressions from the Consumer Health Vocabulary (CHV) [35], used to indicate adverse events, syndromes or conditions. The entries are organized in groups of synonyms. The original resource contains 158,519 entries, many of which are very general (e.g., the words teenager, man, woman). To reduce the ambiguity from the words which are not specific to the domain of drug side effects, we have only used the entries which have been also used as a side effect in a report in FAERS and their synonyms. Each entry in the CHV is assigned to a concept identifier (called a CUI) from UMLS [2] and we use these identifiers as binary features. In addition, for each gazetteer $G$, we compute the following features:

- BIN_$G$: A binary feature, which is set equal to 1 if a tweet contains at least one sequence of tokens matching an entry from gazetteer $G$.
- NUM_TOKENS_$G$: The number of tokens matching entries from gazetteer $G$.
- PRCNT_CHARS_$G$: The fraction of the number of characters in tokens matching entries from gazetteer $G$, relative to the total number of characters in the tweet.

For the example tweet shown in Figure 4, we set the feature BIN_DRUGS equal to 1 because Tylenol matches one entry from the lexicon of drugs. We also set the feature NUM_TOKENS_DRUGS equal to 1 because there is exactly one token marked as drug, and PRCNT_CHARS_DRUGS = 7 / 73 = 0.0959, given that the tweet is 73 characters long.

## 4.5 Sentiment

The sentiment of users as expressed in their tweets is potentially an important indication regarding their experience with pharmaceutical drugs mentioned in tweets. To leverage this information in the classification, we employ a dictionary-based approach, using the resource described by Hansen et al. [9]. Each word in the dictionary is associated with a rating for valence or polarity between $-5$ and $+5$. We only keep dictionary entries with a rating greater than $+2$ or less than $-2$ in order to focus on words expressing strong sentiments. During feature extraction, we assign to each word $w$ in the tweet its valence rating. Next, we aggregate the positive and negative ratings separately. The features we generate are the following:

- $F$_OF_NEGATIVE_PHRASES: For $F \in \{$NUM, SUM, AVG$\}$, we compute the number of tokens with a negative index, their sum and their average, respectively.
- $F$_OF_POSITIVE_PHRASES: For $F \in \{$NUM, SUM, AVG$\}$, we compute the number of tokens with a positive index, their sum and their average, respectively.

For the example tweet shown in Figure 4, there is one positive phrase ("better") and no negative phrases. Hence, we compute three sentiment features, the number of positive phrases, which is equal to 1, the sum of positive phrases, which is equal to 3, the valence rating of "better" and the average, which is equal to 3 as well.

Our sentiment features do not conflate positive and negative valence scores unlike [24].

# 5. RESULTS

In this section, we present a quantitative and qualitative evaluation of our system's classification quality. For our quantitative experiments we split our gold data into three sub-sets, assigning 80% for training, 10% for development and 10% for testing. We arrange the annotated tweets in chronological order and form the training set with the 8,000 earliest tweets. We randomly shuffle the remaining 2,000 tweets and partition them in the development and test sets. We optimize the parameter $C$ and the weight $w^+$ of the positive training examples for the SVM classifier by searching for the values of $C$ and $w^+$ that maximize F1 in the development set. The negative class weight is $w^- = 1$.

Table 1 shows the results of our quantitative evaluation. The first row from the table shows the effectiveness of a baseline classifier, which classifies a tweet as positive, i.e., as mentioning an adverse event if there is a phrase in the text that matches an entry from the Consumer Health Vocabulary (Section 4). We did not compare our automatic classifier directly with the semi-automatic method of [6] because they use a different categorization system, or with the classifiers described in [1] because their approach involves a two-stage classification which is not compatible with our setting.

The second row shows the performance of a classifier learned using only the binary $n$-gram features, e.g., BIN_NGRAM1 and BIN_NGRAM2. This classifier achieves F1=54.9% on the test set and corresponds to an improvement of 51% compared to the baseline in terms of F1 score for the test set, which is statistically significantly better at $p < 0.01$ than the baseline according to the McNemar test [26].

The last row of Table 1 shows the performance of a classifier trained on all surface, gazetteer, POS and sentiment analysis features added to the binary unigram and bigram features. The F1 score increases 3.6% from 0.562 to 0.582 on

**Table 1: Evaluation results for keyword-based classifiers. $*$ indicates significance at $p < 0.01$ w.r.t. the baseline and † indicates significance at $p < 0.01$ w.r.t. the classifier learned from unigrams and bigrams.**

| | | | Development | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | $C$ | $w^+$ | P | R | F1 | P | R | F1 | |
| Baseline | | | 0.269 | 0.683 | 0.386 | 0.267 | 0.705 | 0.387 | |
| BIN_NGRAM1,2 | 0.050 | 8 | 0.636 | 0.504 | 0.562 | 0.560 | 0.540 | 0.549 | $*$ |
| ALL FEATURES | 0.025 | 9 | 0.573 | 0.590 | 0.582 | 0.550 | 0.669 | **0.604** | $*$† |

**Table 2: Confusion matrix.**

| | | Ground truth | | |
|---|---|---|---|---|
| | | True | False | Total |
| Prediction | True | 93 | 46 | 139 |
| | False | 76 | 785 | 861 |
| | Total | 169 | 831 | 1000 |

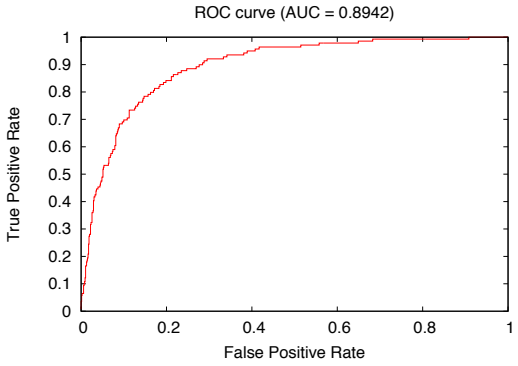


**Figure 5: The ROC curve for the model trained on all features.**

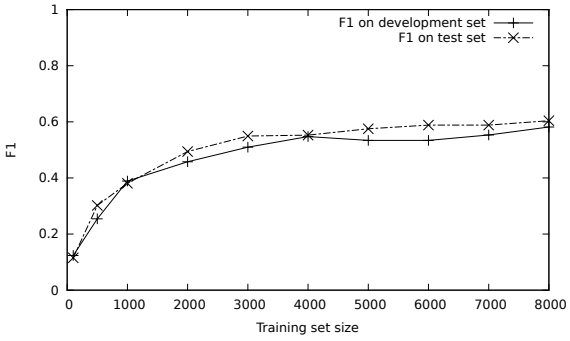


**Figure 6: Learning curve of the classifier.**

the development set. On the test set, we obtain F1=0.604. According to the McNemar test, the classifier using all features is statistically significantly better than both the baseline and the classifier learned from unigrams and bigrams at $p < 0.01$. When we compare the baseline and the model trained on all features, we can see that the latter achieves significantly higher precision with a small loss in recall. This is expected given the inverse relation between precision and recall. Also, since the parameters were set in order to optimize F1 score in the development set, instead of precision and recall, we accept a small loss in recall compared to the important gain in precision. Table 2 shows the confusion matrix of the model learned from all the features evaluated on the test set. Figure 5 shows the ROC curve for the model corresponding to the last row in Table 1.

Table 3 shows examples from the output of our system. The examples in the first 5 rows are correctly classified as mentioning drug side effects. The sixth row is a true negative because it is implied that a health professional is giving feedback about the side effect of a drug on patients. The last 3 examples correspond to misclassified tweets. The example in the 7th row refers to fatigue using the phrase "I seriously never have any energy." Finally, the last two examples are comments of users about the efficacy of the mentioned drugs.

Figure 6 shows how the F1 score of the classifier changes as we train it with 100, 500, 1,000 examples, respectively, and increase the training set size to 8,000 examples in increments of 1,000. We can see that the F1 score on both the development and the test set increases as we use more training examples. As the learning curve in Figure 6 shows, the current amount of training data provides evidence for the learner to the effect that adding more data points for training has only a diminished return (in terms of any potential F1 score increase for any further annotation cost from then onwards), i.e., the performance is increasing only slowly as a plateau has nearly been reached. Moreover, we observe that the F1 score obtained in the test set is higher than the F1 score in the development set. This may be attributed to the development and test sets being similar, since they were collected during the same time period.

## 6. SIGNAL YIELD ANALYSIS

Our second research question, which is pertinent to the use of social media, and Twitter in particular, for mining of adverse events, is related to the estimation of the volume of data that are relevant to the task. Given that there are 500 million tweets posted daily, it is important to understand how many of all these tweets are actually discussing adverse events. To answer this question, we ran a large-scale data collection from Twitter, spanning from July 9 to September 4, 2014. Using Twitter's search API, we repetitively submitted queries consisting of a drug name from a list of more than 2,600 drug names, and requested the most recent tweets that matched each query. We have collected 1.5 million tweets in total.

Then, we computed the volume of tweets corresponding to the output of the topic filter, the volume filter and the binary classifier. Figure 7 describes the sequence of filters and the classifier in the layered architecture of our system, responsible for reducing the data volume from the 500 million tweets per day to an average of 721 tweets per day, in which Twitter users mention adverse events.

We also validated our approach by accessing 100% of tweets using the commercial aggregator datasift.com. The

**Table 3: Example system output: correct (top) and incorrect (bottom).**

| | Tweet text |
|---|---|
| TP | wellbutrin doesnt even work for me it just makes me really anxious idk why im still taking it |
| TP | Took some ibuprofen that has me so drowsy |
| TP | Insomnia and heart palpitations due to prednisone. Takteng mga side effects 'to. /wrist |
| TP | This Vicodin is making me feel like a tweaker because I'm so itchy! |
| TP | guys i took an ibuprofen 2 days ago and i still have heart palpitations |
| TN | I once gave a treatment to some patients with glutathione which I knew later clinic use it. Most of them have nausea for 5 mins after inj. |
| FN | I seriously never have any energy thanks accutane lol @probs_accutane all I want to do is sleep |
| FP | My mouth taste like 1200 Mg's of ibuprofen yet my head still hurts and I'm still feeling dizzy. Wtf |
| FP | @ash_hein they gave me Tylenol 3s & yea kinda my mouth still hurts a little & I'm still swollen |

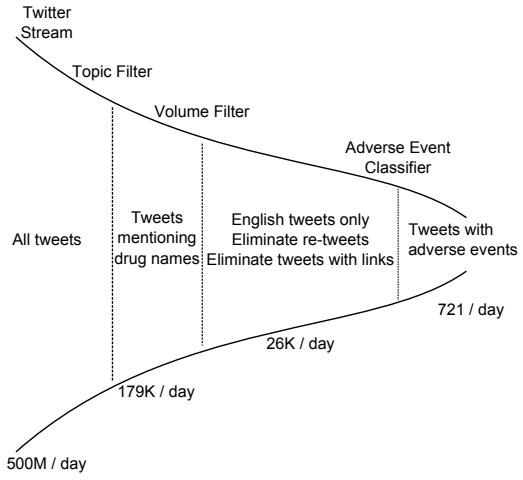


**Figure 7: The funnel: the layered architecture of our system effectively reduces data volume.**

data collection was performed with a subset of 187 drugs for 30 days, starting on September 4, 2014. We have collected 290,000 tweets out of which an average of 405 tweets per day were marked as positive by the classifier. The yield of our system is similar in both settings where we have used Twitter's Search API and the full firehose.

The yield of the system depends on both the list of drug names used to collect data, as well as the annotated data used to train the classifier. However, we do not expect that the order of magnitude would be significantly different if we used a different classifier or a training set.

The main implications of our experiment are that the volume filter is performing adequately, i.e., it is capable of reducing the volume to an amount of tweets that is manageable in a fully automatic processing regime. The yield of positives automatically mined by the system from the full firehose is large enough to allow for some further processing and statistical inferences. At the same time, the order of magnitude of the adverse event (English) tweets mined by our platform per day remains small enough so that a single human analyst (e.g. at a regulator or a pharmaceutical company) can reasonably be expected to be able to review them on a daily basis without running up a back-log.

# 7. TOPIC ANALYSIS OF TWEETS

For the training and evaluation of the adverse event classifier, we have relied on a dataset of tweets that mention at least one pharmaceutical drug. During the annotation process, we have focused on identifying these tweets that discuss an adverse drug event. However, we have also observed that there are other topics that are being discussed in the tweets that mention pharmaceutical drugs.

To answer our third research question, we extended the binary annotation scheme to 10 classes, which are pertinent to the topic of pharmacovigilance and the pharmaceutical industry in general. Table 4 shows the description of each class with a representative annotated tweet. This detailed annotation scheme was chosen because we wanted to learn more about the data, including identifying additional opportunities such as building ensemble classifiers over predictors for other classes such as "pharmacological news" or "efficacy," which could be potential confusion classes with our target class. One annotator has applied the extended annotation scheme to the 1,000 of the test set. A tweet may belong to more than one class.

Figure 8 shows the frequency of annotations for the 1,000 tweets of the test set. Not taking into account Class 0, we can see that Class 4 is the most frequent one, followed by Class 5, which corresponds to the topic of drug efficacy. Two other classes that appear with a rate of 1% are Class 2 and Class 3, indicating feedback from patients and professionals, respectively.

The sample we have annotated with the extended scheme contains only very few tweets which have been annotated as Class 1, 2, 7 and 9 and no tweet annotated as Class 8. There are three main reasons for this observation. First, we have not analyzed tweets containing hyperlinks, because such tweets are dropped by the Volume Filter (Section 3.1). When online vendors use Twitter to promote their business, their tweets typically contain a hyperlink to their site to drive traffic and increase sales. Second, the topics that are relevant to classes 6-9 are sensitive to the time range of data collection. For example, if there is no major regulatory event or a deal between pharmaceutical companies during the period for which we have collected tweets, then the number of tweets belonging to Class 6 is expected to be low. Third, we have collected data based on a list of keywords containing names of pharmaceutical drugs, while tweets that are likely to belong to classes 6-9 are about the pharmaceutical companies and are more likely to mention the company name instead of one of its drugs.

# 8. RELATED WORK

Recently, the automatic extraction of adverse events from social media has been investigated as an additional source of information to identify adverse drug events. Due to space limitations, we review the works that are most related to this one. Karimi et al. [12] and Sarker et al. [23] provide extensive reviews of related work about pharmacovigilance

**Table 4: The 10-class annotation system used for data labeling with example tweets for each class.**

| Label | Description | Example |
|---|---|---|
| 1 | Online Vendors | @tyrus_ we do have Nexium 20mg for Ksh 100 and 40 mg for 150. Come check us out at Danchem pharmacy on Duruma rd |
| 2 | Patient feedback: availability, cost | 2,000 for my insulin on my new insurance. Now I cant afford to even get better ! Why keep trying ? |
| 3 | Professional feedback (e.g. comment from a medic) | Tamiflu/Oseltamivir is a neuraminidase inhibitor & blocks viral release, so ideally should be taken within a day or 3 of onset of symptoms |
| 4 | Adverse event | this Vicodin is making my face itchy smh. |
| 5 | Efficacy | Why isn't this Vicodin kicking in ??? |
| 6 | Deals, regulatory, patent-related | EU Pharma Law (C-269/14 P): #ECJ confirms on appeal the suspension of marketing and withdrawal of #Clopidogrel based medicinal products |
| 7 | Clinical trials | Oramed to Announce Results of its Phase 2a Clinical Trial for Oral Insulin on January 30th at the Tel Aviv Stock Exchange ... |
| 8 | Financial aspects of pharmaceutic industry | Pfizer aims to keep one-third of Lipitor pie |
| 9 | Pharmaceutic news | Testosterone therapy is a gigantic test being done on US men @CBSEveningNews And that means ad money for CBS, @ABCWorldNews @nbcnightlynews |
| 0 | Everything else | #LastDance is so intense I might need a Tylenol pm after this |

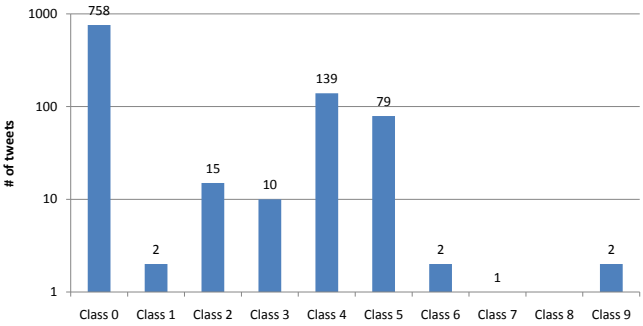


**Figure 8: Number of tweets assigned to each of the ten classes in the extended annotation scheme. The scale of the y axis is logarithmic.**

from online social media.

## 8.1 Mining Health-related Web forums

Leaman et al. [15] present one of the first works to attempt extracting ADRs from online forums. They experiment with a set of four drugs, obtaining F1=73.9%. They identify ADRs with side effect dictionaries, as well as a list of manually selected set of colloquial phrases, using sliding windows and Jaro-Winkler string similarity for approximate string matching.

Nikfarjam and Gonzalez [20] employ association mining and the Apriori algorithm to identify ADRs from user comments in the health-related forum DailyStrength[7]. They identified rules where the condition consists of $n$-grams of POS tags and the consequence is an ADR. The conditions of the identified rules are then converted to patterns which match the text of user posts. The obtained P=70%, R=66% and F1=68% on a corpus of 3,600 annotated posts.

Chee, Berlin and Schatz [4] analyze messages posted on Yahoo! forums about health issues to extract ADRs and identify drugs, which have undergone label changes or regulatory action. The authors employ bagging of SVM classifiers with Radial Basis Function (RBF) kernel and Naïve Bayes classifiers. The classifiers are not explicitly evaluated. Instead, they are used to identify drugs as being candidates to add to a watchlist if they are consistently classified as false positives. The ranking of drugs is based on an ad-hoc formula with manually set weights.

Liu, Li and Seneff [16] study the problem of discovering ADRs caused by the use of statin drugs. Based on data extracted from online forums, they develop a taxonomy of side effects, and they use log-likelihood ratios to identify the side effects that are significantly associated with statins.

Yang et al. [32] employ association mining and metrics such as lift, leverage and Proportional Reporting Ratio [5], to rank ADRs based on their occurrences in posts of users in health-related forums. The ranking of ADRs is compared with ADRs reported by FDA. The matching of ADRs in text is performed using the Consumer Health Vocabulary (CHV) and a sliding window of n-grams for approximately matching the lexicon terms. The experimental results involve only 10 drugs and indicate that PRR is the most effective measure.

Wu, Fang and Stanhope [30] describe an early warning system for ADRs, which uses a knowledge base of more than 4,000 drugs and 66 side effects per drug on average, created from SIDER [13] and online health forums. Mutual information between side effects is used to identify related side effects. For each message, a score is computed based on the occurrences of a side effect as well as related ones. If the computed score is greater than a threshold, then it is assumed that the message is discussing an ADR.

Yates, Goharian and Frieder [33] extract drug side effects from consumer posts in health forums related to breast cancer using a Conditional Random Field (CRF) model over features comprising linguistic dependency relations, terms, POS tags and thesaurus matches (P=61%, R=32%).

Metke-Jimenez, Karimi and Paris [19] provide a detailed comparison of text preprocessing methods applied to the extraction of adverse events, diseases and drug mentions from the online forum AskAPatient. Their evaluation results show that stemming is not helpful in any of the test settings, the most effective tokenizer is based on the Unicode Text Segmentation algorithm[8], and also that the most effective vocabulary overall is the CHV, which outperforms technical medical vocabularies.

Metke-Jimenez and Karimi [18] use a CRF tagger to identify adverse drug reactions in health-related forums, using

[7]http://www.dailystrength.org/

[8]http://unicode.org/reports/tr29/

a set of 1,250 reviews for 12 drugs. They train a CRF tagger for identifying spans of adverse drug reactions in drug reviews (F1=60.2%), and use a terminology server to map spans to entries in an ontology of medical concepts (F1=50.6%).

Nikfarjam et al. [21] employ a CRF tagger to extract mentions of adverse drug reactions from posts in the DailyStrength forum (4,720 reviews used for training and 1,559 reviews used for testing) and Twitter (1,340 tweets used for training and 444 used for testing). The features of the CRF tagger include the words and their tokens, as well as cluster features learned from distributed word representations. They report F1=72.1% for Twitter data.

Karimi et al. [11] present a system for mining online health forums to identify adverse drug events. They focus on finding mentions of previously unknown side effects by comparing the list of extracted side effects with existing lists of known side effects for drugs.

## 8.2 Mining Twitter

Bian, Topaloglu and Yu [1] extract adverse events using a SVM with RBF kernel and a set of features (bag of words, MetaMap semantic classes, pronouns, numeric features). Using a set of five drugs at a clinical trial stage, with a total of 239 examples, they obtain accuracy A=74%.

In perhaps the most closely related work to ours, Sarker and Gonzalez [24] use data from three different sources, including a dataset of tweets [7], and develop automatic classifiers to detect adverse events for drugs. They report F1=53.8% when training a SVM with RBF kernel on the dataset of tweets. When the training data is augmented with other datasets from medical case reports and DailyStrength, they report F1=59.7%. However, they do not attempt to estimate the signal yield.

Freifeld et al. [6] present a comparison study between adverse events found on Twitter and FAERS. Starting out with 6.9 million tweets collected over a seven-month period using Twitter's Search API, they used a set of 23 drug names and a list of symptoms to reduce that data to a subset of 60,000 tentative tweets, which were examined by humans in a manual procedure that resulted in 4,401 positive messages. Spearman rank correlation between the profiles of ADEs reported on Twitter and FAERS was found to be $\rho = 0.75$.

Yates, Goharian and Frieder [34] use Labeled LDA, a Naive Bayes classifier and a CRF tagger to identify adverse drug reactions in a health-related forum and Twitter. The CRF tagger achieves consistently higher precision compared to the other approaches. However, in the case of the health forum data, the baseline from [33] achieves higher F1 score. The ground-truth used for Twitter data is based on a list of known side effects of drugs listed in SIDER and is used to compare the relative ranking of the different methods with the resulting ranking of methods from the experiments with the health forum data.

Finally, Sarker, Nikfarjam and Gonzalez [25] present the results from a shared task with the aim to evaluate the performance of systems in the tasks of classifying tweets mentioning adverse drug events, information extraction of adverse drug events from tweets, and the normalization of user mentions of adverse events to standardized ontologies.

All this previous work pertains to various aspects of our system presented here. Most of the studies use health-related forums and studies using Twitter have been carried out either at a small scale, involving manual steps, or finding only known side effects. We are not aware of any previous attempts to quantify the yield of an ADE classifier for tweets, or to analyze topics discussed in tweets mentioning pharmaceutical drugs, other than side effects.

# 9. CONCLUSIONS

We have presented a system for large-scale pharmacovigilance support. We have addressed the problem of adverse event extraction from tweets by training and evaluating a supervised binary classifier based on SVM on 10,000 manually annotated tweets. For each tweet, we extract a set of features based on words and keywords, surface features, a list of gazetteers, POS tags and sentiment analysis. To answer our first research question, we have evaluated the adverse event classifier. The best model using all the features obtained F1=60.4%, which is statistically significantly better compared to the baseline (F1=38.7%). Based on a large dataset of 1.5 million tweets, we have answered our second research question and observed that the classifier yields on average 721 new tweets per day. While this number depends on the number of covered drugs and the classifier, it indicates the order of magnitude for the volume of tweets discussing adverse events on Twitter. To answer our third research question, we have investigated what other topics are discussed in tweets mentioning pharmaceutical drugs, and we have found that efficacy of drugs is the second most frequent class after adverse events. It is important to note that the prevalence of the classes depends on the data collection methodology. Hence, when developing a system for a specific class, the data collection must consider the specific characteristics of the data in the class for optimal performance.

Our results indicate that social media can be used in the ecosystem of regulators and pharmaceutical companies to obtain feedback from the users of social media and complement existing communication channels.

There are several avenues for future investigations. Currently, we treat each message as an independent post to reduce processing complexity. In reality, there are patient conversations happening on Twitter as well, where mentions of a drug and a related adverse event occur in separate tweets. This information, as well as follow relationships between users could be useful to provide additional context and improve the system, but at a higher processing cost.

Most tweets mention a single drug, but there is a number of tweets mentioning two or more drug names. In future work, we plan to use our existing platform to mine and quantify the extent to which Twitter contains mentions of Drug-Drug Interactions (DDI). Moreover, our existing platform permits experimentation with the relationship of branded pharmaceutical drugs and generics. For example, the platform can be used to compare whether conversations about one drug occur significantly more often than another.

Finally, this paper reported results from a binary classifier that is trained to identify adverse events; user-generated data like tweets also permits other, related applications such as studying the efficacy of a drug. To explore these further, we can reuse our ten class annotation scheme.


## Acknowledgments

The authors would like to thank Miriam Bayés Genís for helping with the manual annotation and Khalid Al-Kofahi

for feedback and support. This paper received financial support by Thomson Reuters Global Resources.

# 10. REFERENCES


[1] J. Bian, U. Topaloglu, and F. Yu. Towards Large-scale Twitter Mining for Drug-related Adverse Events. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, SHB '12, pages 25–32, 2012.

[2] K.E. Campbell, D.E. Oliver, and E.H. Shortliffe. The Unified Medical Language System: Toward a Collaborative Approach for Solving Terminological Problems. *J. Am. Med. Inform. Assoc.*, 5:12–16, 1998.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.

[4] B.W. Chee, R. Berlin, and B. Schatz. Predicting Adverse Drug Events from Personal Health Messages. In *AMIA Annual Symposium Proceedings*, pages 217–226, 2011.

[5] S.J.W. Evans, P.C. Waller, and S. Davis. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10:483–486, 2001.

[6] C.C. Freifeld, J.S. Brownstein, C.M. Menone, W. Bao, R. Filice, T. Kass-Hout, and N. Dasgupta. Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter. *Drug Safety*, 37(5):343–350, 2014.

[7] R. Ginn, P. Pimpalkhute, A. Nikfarjam, A. Pakti, K. O'Connor, A. Sarker, K. Smith, and G. Gonzalez. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. In *Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, pages 1–8, 2014.

[8] K.L. Gwet. *Handbook of Inter-Rater Reliability*. Advanced Analytics LLC, 4th edition, 2014.

[9] L.K. Hansen, A. Arvidsson, F.A. Nielsen, E. Colleoni, and M. Etter. Good Friends, Bad News - Affect and Virality in Twitter. In *Future Information Technology*, volume 185 of *CCIS*, pages 34–43. 2011.

[10] L. Hazell and S.A.W. Shakir. Under-Reporting of Adverse Drug Reactions. *Drug Safety*, 29(5):385–396, 2006.

[11] S. Karimi, A. Metke-Jimenez, and A. Nguyen. CADEminer: A system for mining consumer reports on adverse drug side effects. In *Proceedings of the 8th Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '15, pages 47–50, 2015.

[12] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris. Text and data mining techniques in adverse drug reaction detection. *ACM Comput. Surv.*, 47(4):56:1–56:39, 2015.

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

[14] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[15] 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 2010 Workshop on Biomedical Natural Language Processing*, pages 117–125, 2010.

[16] J. Liu, A. Li, and S. Seneff. Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs. In *Proceedings of the 1st International Conference on Advances in Information Mining and Management*, pages 91–96, 2011.

[17] M. Lui and T. Baldwin. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 25–30, 2012.

[18] A. Metke-Jimenez and S. Karimi. Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms. http://arxiv.org/abs/1504.06936, 2015.

[19] A. Metke-Jimenez, S. Karimi, and C. Paris. Evaluation of Text-Processing Algorithms for Adverse Drug Event Extraction from Social Media. In *Proceedings of the International Workshop on Social Media Retrieval and Analysis*, 2014.

[20] A. Nikfarjam and G.H. Gonzalez. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings*, pages 1019–1026, 2011.

[21] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J. Am. Med. Inform. Assoc.*, 22:671–681, 2015.

[22] M. Paul and M. Dredze. A Model for Mining Public Health Topics from Twitter. Technical report, Department of Computer Science, John Hopkins University, 2011.

[23] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. Utilizing social media data for pharmacovigilance. *Journal of Biomedical Informatics*, 54:202–212, 2015.

[24] A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207, 2015.

[25] A. Sarker, A. Nikfarjam, and G. Gonzalez. Social media mining shared task workshop. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 581–592, 2015.

[26] S. Siegel and N.J. Castellan Jr. *Nonparametric statistics for the behavioral sciences*. McGraw–Hill, Inc., 2nd edition, 1988.

[27] J.S. Uebersax. Diversity of decision-making models and the measurement of interrated agreement. *Phychological Bulletin*, 101(1):140–146, 1987.

[28] S. Wang and C.D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 90–94, 2012.

[29] World Health Organization. The Importance of Pharmacovigilance. http://bit.ly/1RI8HcE, 2002.

[30] H. Wu, H. Fang, and S.J. Stanhope. An Early Warning System for Unrecognized Drug Side Effects Discovery. In *Proceedings of the 21st International Conference on World Wide Web companion*, pages 437–440, 2012.

[31] X. Yan, J. Guo, S. Liu, X.-Q. Cheng, and Y. Wang. Clustering Short Text Using Ncut-weighted Non-negative Matrix Factorization. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2259–2262, 2012.

[32] C.C. Yang, H. Yang, L. Jiang, and M. Zhang. Social Media Mining for Drug Safety Signal Detection. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, pages 33–40, 2012.

[33] A. Yates, N. Goharian, and O. Frieder. Extracting Adverse Drug Reactions from Forum Posts and Linking them to Drugs. In *Proceedings of the ACM SIGIR Workshop on Health Search & Discovery: Helping Users & Advancing Medicine*, pages 55–58, 2013.

[34] A. Yates, N. Goharian, and O. Frieder. Extracting adverse drug reactions from social media. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2460–2467, 2015.

[35] Q.T. Zeng, J. Crowell, G. Divita, L. Roth, and A.C. Browne. Identifying Consumer-Friendly Display (CFD) Names for Health Concepts. In *AMIA Annual Symposium Proceedings*, pages 859–863, 2005.