Mining Adverse Drug Reaction Signals from Social Media: Going Beyond Extraction

Apurv Patki[1], Abeed Sarker[2], Pranoti Pimpalkhute[1], Azadeh Nikfarjam[2], Rachel Ginn[2], Karen O'Connor[2], Karen Smith[2], Graciela Gonzalez[2]

1. Dept. of Computer Science, Arizona State University, 2. Dept. of Biomedical Informatics, Arizona State University

ABSTRACT

The recent popularity of health related social networks has enabled users to communicate about drugs, treatments and other health related issues over the Internet, making it a rich resource for monitoring drugs after they hit the market. In this paper we explore a novel probabilistic model for drug categorization using a two-step approach. We first classify whether a comment includes a mention of an adverse drug reaction, and then infer whether the combined comments for the drug (its social media discourse) indicate a potential red flag, an inordinate incidence of adverse reactions. The best classifier for the classification of ADR assertive comments reaches an accuracy of 82% with the ADR class F-score of 0.652, which is an important step forward in extracting actual mentions. Utilizing the comments to infer whether the drug is behaving in an abnormal manner proved a more challenging problem, and our results are marginal but promising on this first attempt.

1 INTRODUCTION

Research has shown that adverse drug reactions (ADR) are associated with severe health and financial consequences: with deaths and hospitalizations numbering in millions, and associated costs of about seventy-five billion dollars annually (Harpaz *et al*., 2012). Detection of adverse reactions associated with drugs once they hit the market is the focus of *pharmacovigilance*, "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug problem" (World Health Organization, 2013).

The rapid growth of electronically available health related information (be it in electronic medical records or social media) plus the advances in Natural Language Processing (NLP) and machine learning algorithms present a unique opportunity to massively mine data for the presence of ADR mentions. Prior work has focused on automatic extraction of ADR mentions from electronic medical records (Aramaki, Miura, & Tonoike, 2010) and from user comments in social media (Nikfarjam & Gonzalez, 2011). However, the question remains: how can these mentions be used in *pharmacovigilance*, for raising a "red flag" when needed?

Health-related social networking sites are more popular than ever, and are generally accepted as a viable platform to discuss health-related experiences, including symptoms and treatments for different diseases, as well as their side effects. Because of the costs associated with post-marketing ADRs caused by drugs, and the large volume of user posted information available in social media, there is a strong motivation for systems that can automatically monitor social media sites and generate signals when adverse reactions frequently occur for specific drugs.

We focus this study on data from one such social network, DailyStrength[1]. According to the survey carried out by Comscore[2] in September 2007, DailyStrength observed 14,000 average daily visitors, spending about 82 minutes on average and each visiting about 145 pages. In this paper, we attempt to address the question of whether it is possible to use the aggregated set of extracted mentions of adverse reactions for a prescription drug to generate a signal, a "red flag" on the map of pharmacovigilance.

1.1 Intents and Contributions

Our primary intent is to explore the possibility of using social media data to identify ADR mentions and to identify potentially harmful drugs through the automatic analysis of user comments. More specifically, our intents in this research are as follows:

(i) Develop automatic classification techniques to identify user comments expressing ADRs from health-related social media data; and

(ii) Assess if the set of probabilities associated with automatically classified user comments expressing ADRs can be utilized to categorize the drug for which they are posted.

We model the problem more broadly than that of detecting a specific unknown adverse reaction. We first try to detect the general discourse of the discussions in social media for a given drug by analyzing individual comments and classifying

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

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

them. Based on the automatic classifications, we explore approaches by which the observed discourse for a drug can be classified as *normal* (what could be expected of any drug that does not pose a serious threat) or *blackbox* candidate (which might point to evidence of adverse reactions). The contributions we make in this paper are as follows:

(i) We discuss how information about ADRs is distributed in social media postings, and potential approaches by which it can be harnessed for use in pharmacovigilance.

(ii) We show that annotated corpora obtained from social media data and specifically annotated for the detection of ADRs can play an important role in pharmacovigilance.

(iii) We present and compare automatic, supervised binary classification approaches that can be used to identify individual comments mentioning ADRs in social media postings. We also discuss possible ways in which automatic classification accuracies can be improved when applied to imbalanced data sets, which is a common obstacle when mining data from social media.

(iv) We present a discussion of possible approaches by which the probabilities assigned by automatic supervised classifiers can be combined to distinguish between *normal vs. blackbox* discourse for the drug.

The rest of the paper is organized as follows: In Section 2, we provide an overview of the related work in this field; we present our annotated corpus, methods, and results in Section 3; we provide a discussion of our findings in Section 4, along with our plans for future explorations; and we conclude the paper in Section 5.

2 RELATED WORK

Most of the previous text mining research related to pharmacovigilance is focused on electronic health records (Aramaki *et al.*, 2010; Friedman, 2009; Wang *et al.*, 2009), and medical case reports (Gurulingappa, Rajput, & Toldo, 2012; Toldo, Bhattacharya, & Gurulingappa, 2012). Harpaz *et al.* (2012) provide a thorough survey on the existing approaches for post-marketing pharmacovigilance, exploring various resources such as electronic health records, spontaneous adverse drug reporting systems and biomedical literature. Social media was relatively unexplored for this purpose until recently. Leaman *et al.* (2010) analyzed user comments in social media and demonstrated that the comments contain extractable drug safety information. The authors used a hybrid lexicon and rule-based system for ADR concept extraction. Nikfarjam & Gonzalez (2011) proposed a pattern-based technique based on association rule mining, which extracts ADR mentions based on the language patterns used by patients in social media for expressing ADRs. In a recent study Yates & Goharian (2013) analyzed the value of user comments in revealing the unknown adverse effects by evaluating the extracted ADRs against the SIDER database[3] which contains information about the known adverse effects. There are similar studies for automatic ADR mention extraction, targeting online patient discussions (Yates & Goharian, 2013; Benton *et al.*, 2011; Sampathkumar, Luo, & Chen, 2012). While these techniques can be used to extract ADR mentions from the available online user contents, our task only requires a binary decision about the comment being ADR or NoADR.

Chee et al. (2011) classified user posts on online groups to predict the candidate FDA watchlist drugs for further investigation with regards to drug safety. They used an ensemble based classification technique to identify drugs that are likely to be in the watchlist category. Our work is different in two ways, first, our dataset is from health related social network, which generally contains unstructured sentences, incorrect spellings, and more informal language compared to electronic health records. Secondly, we hypothesize that a drug can be classified as watchlist (we refer to these as *blackbox*) or *normal* based on the amount of adverse events that are reported about the drug.

3 DATA COLLECTION AND ANNOTATION

3.1 Drug name Identification

The first step in our data collection process involved the identification of a set of drugs to study, followed by the collection of user comments associated with each drug name. To maximize our ability to find relevant comments, we focused on two criteria: (i) drugs prescribed for chronic diseases and conditions that we might expect to be commonly commented upon, and (ii) prevalence of drug use. For the first criterion, we selected drugs used to treat chronic conditions such as type 2 diabetes mellitus, coronary vascular disease, hypertension, asthma, chronic obstructive pulmonary disease, osteoporosis, Alzheimer's disease, overactive bladder, and nicotine addiction. To select medications that have a relatively high prevalence of use and thus exposure, we selected drugs from the IMS Health's Top 100 drugs by volume for 2013[4]. Medication categories of interest that we identified in the Top 100, which were not in our chronic disease list, included attention deficit hyperactivity disorder stimulants, anti-retrovirals, biologics, thyroid hormones, influenza treatment and vaccine, oral contraceptives, oral anticoagulants, anti-depressants, erectile dysfunction and non-steroidal anti-inflammatory drugs. Next, we categorized

[3] http://sideeffects.embl.de/

[4] http://www.imshealth.com/portal/site/imshealth

the selected drugs into three classes: *normal*, boxed warnings (*blackbox*), and *withdrawn*. These categorizations were based on the manufacturers' package inserts and FDA information. Drugs in the normal category had no black box warning or history of withdrawal from the market; however, they could have associated warnings and precautions. Blackbox drugs had associated FDA-issued *blackbox* warnings due to identified serious or life-threatening safety concerns. Finally, the withdrawn category included drugs that were withdrawn from the market in any country, or for any length of time. For the research described in this paper, we only target the automatic categorization of normal and blackbox drugs.

3.2 Comment collection and annotation

We obtained comments associated with each drug from DailyStrength, a health related social network where people share their personal knowledge and experiences regarding diseases and/or treatments, among other things. For the drugs selected for this study, we collected 20,486 comments (normal: 10,399, blackbox: 7,327, and withdrawn: 2,760) from the review pages. The user comments are not evenly distributed among drugs, and some drugs have very few associated comments. Each treatment, or drug, has a specific review page.

A subset of the user comments (10,617 in total) was annotated by two domain experts under the guidance of a pharmacology expert. The comments are annotated for adverse drug effects, indication, beneficial effects, and other mentions. For annotation, we defined an adverse drug effect as "an undesired effect of the drug experienced by the patient." This included mentions where a patient expressed the notion that a drug worsened their condition. An indication was defined as "the sign, symptom, syndrome, or disease that is the reason or the purpose for the patient taking the drug or is the desired primary effect of the drug. Additionally, the indication is what the patient, prescriber, etc. believes is the main purpose of the drug." Beneficial effects as defined for this study are "an unexpected effect of the drug that positively impacted the patient." The annotated spans were mapped to UMLS concept IDs found in the lexicon.

Our lexicon (Ginn *et al.*, 2014) was derived from the lexicon used by Leaman *et al.* (2010), which includes terms and concepts from four resources. The COSTART vocabulary created by the U.S. Food and Drug Administration for post-marketing surveillance of ADRs (a subset of the UMLS Metathesaurus), which contains 3,787 concepts; the SIDER side effect resource – which contains 888 drugs linked with 1,450 adverse reaction terms extracted from pharmaceutical insert literature, and the Canada Drug Adverse Reaction Database, or MedEffect[5], which contains associations between 10,192 drugs and 3,279 adverse reactions. These resources provided both the concept names and the UMLS IDs. The lexicon was manually reduced by grouping terms with similar meanings, for example "appetite exaggerated," and "appetite increased". We added additional terms from SIDER II (Kuhn *et al.*, 2010) and the Consumer Health Vocabulary (CHV) (Zeng-Treitler *et al.*, 2008), which includes more colloquialisms.

An initial set of comments was annotated by each annotator. Discussions about these annotations were held with the annotators and the pharmacology expert. From these discussions, annotation rules were created and this formed the annotation guidelines document that was followed for the remaining annotations. An example rule describes scope of the 'discontinuation' annotations in the 'other' category: they should span the minimal terms needed to communicate that the treatment was stopped, but not including policy changes (like taken off the market). The pharmacology expert also reviewed the annotations and created the gold standard. Cohen's Kappa (Carletta, 1996) value for the inter annotator agreement (IAA) is 0.67, which represents 'significant agreement' between the two annotators. Since this paper analyzed only the binary presence of an ADR (even though other annotations are available), the Kappa applied to the binary presence of an ADR within each post.

4 METHODS

4.1 Experiments

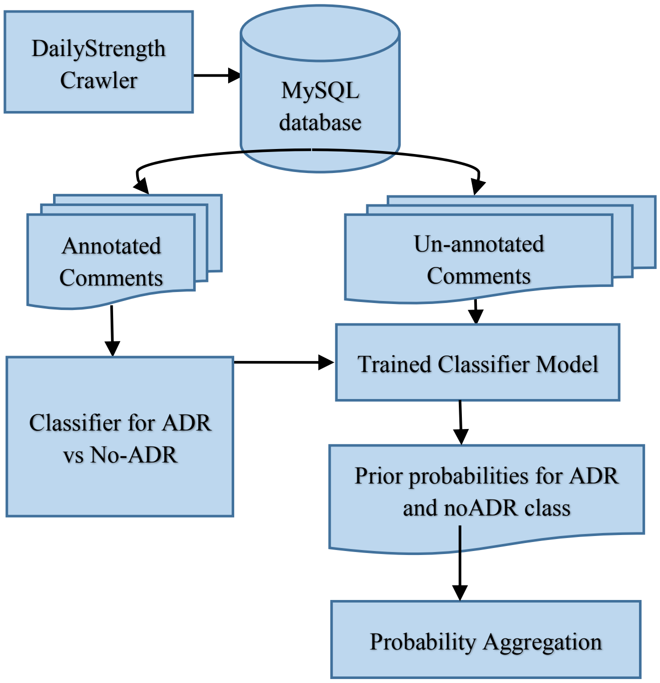

Figure 1. Flowchart illustrating our two-step drug classification process.

[5] http://hc-sc.gc.ca/dhp-mps/medeff/index-eng.php

As explained earlier, the intent of our research is to explore if drugs can be classified automatically into normal *vs.* blackbox categories utilizing the comments associated with them. Therefore, from our data set, we only use the annotations that indicate whether a comment presents an ADR or not (ADR *vs.* noADR). In the first step of our two-step approach, we use these annotations to automatically classify user comments. Our intuition is that drugs within the blackbox categories should have greater incidence of adverse reactions associated with them. In the second step, we combine the probabilities of the classified comments to automatically predict if a drug should be categorized as *normal* or *blackbox*. In the following subsections, we detail the approaches for these two steps. Figure 1 graphically illustrates our approach.

4.1.1 Binary Classification

For the binary classification of comments into ADR and noADR categories, we use two supervised machine learning algorithms: Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM). MNB is a common and simple supervised learning algorithm, which is often used for comparisons. SVMs have been shown to perform particularly well for supervised text classification due to their capability to deal with high dimensional feature spaces, dense concept vectors, and sparse instance vectors. Our data for this part of the experiments consists of 10,617 manually annotated user comments from DailyStrength, belonging to all the three drug categories mentioned previously. 2,513 (23.7%) of these comments belong to the ADR category, while 8,104 (76.3%) comments belong to the noADR category. As the numbers suggest, the data is imbalanced with an ADR to noADR ratio of 1:3.2. We attempt to address this imbalance using a cost sensitive classification scheme described later. We use some simple NLP techniques outlined below to preprocess the comments and extract features from them.

Pre-processing. We preprocess the comment texts by lowercasing the characters and stemming all the terms using the Porter stemmer[6].

N-grams. Our first feature set consists of word n-grams of the comments. We use 1-, 2-, and 3-grams as features.

Synset Expansions. It has been shown in past research that certain terms play an important role in determining the polarities of sentences (Sarker, Molla, & Paris, 2013). Since the binary classification of comments is similar to automatic, sentence-level polarity classification, we incorporate this feature into our classification task. For each adjective, noun or verb in a sentence, we use WordNet[7] to identify the synonyms of that term. We then add all the synonymous terms in a bag-of-words manner, attached with the 'SYN' tag, as features.

Change Phrases. We use the Change Phrases features proposed by Niu *et al.*, (2005). The intuition behind this feature set is that a sentence represents positive information or negative information can often be signaled by how a change happens: if a bad thing (*e.g.*, headache) was reduced, then it is a positive outcome; if a bad thing was increased, then the outcome is negative. This feature set attempts to capture cases when a good/bad thing is increased/decreased. We first collected the four groups of good, bad, more (increase), and less (decrease) words used by Sarker *et al.* (2013). This feature set has four features: MORE-GOOD, MORE-BAD, LESS-GOOD, and LESS-BAD. We applied the same approach as Niu *et al.* (2005) (*i.e.*, window of four terms) to extract this feature. The features are represented using a binary vector with 1 indicating the presence of a feature and 0 indicating absence.

Sentiword. Our inspection of the data suggests that comments associated with ADRs generally present negative sentiment. For this feature, we attempt to incorporate a score that represents the general sentiment of a comment (as the normalized sum of all the terms in the comment). Each word in a comment is assigned a score and the overall score assigned to the comment is equal to the sum of all the individual term sentiment scores, normalized by the length of the sentence in words. To obtain a score for each term, we use the lexicon proposed by (Guerini, Gatti, & Turchi, 2013)[8]. The overall score a sentence receives is therefore a floating point number with the range [-1:1].

4.1.2 Binary Classification Results

We train two machine learning classifiers using the features mentioned above: MNB and SVM. For both the classifiers, we use the implementations provided by the machine learning toolbox Weka[9]. We assess the performance of the two approaches via 10-fold cross validation over our annotated data set of 7,693 comments. For the SVM classifier, we use a polynomial kernel and the complexity parameter = 1.0. To obtain probability estimates for the predictions by this classifier, we fit a logistic regression model to the outputs of the SVMs. For both classifiers, at each fold of the 10-fold cross validation, we reduce the feature space by only keeping useful features. For this, we use the information gain attribute evaluation for each individual fold, and we only keep the most informative 1,500

[6] We used the stemmer provided by the NLTK toolkit: http://www.nltk.org/

[7] http://wordnet.princeton.edu/

[8] The lexicon is available for download at: https://hlt.fbk.eu/technologies/sentiwords

[9] Available from: http://www.cs.waikato.ac.nz/ml/weka/

attributes. To address the issue of data imbalance, we apply a cost sensitive classification strategy. Using this approach, the training instances are reweighted according to a total cost assigned to each class. To assign costs, we apply an explicit cost matrix and the cost assigned to each class is equal to its ratio in the data set (*i.e.,* 1 and 3.2 for the ADR and noADR classes, respectively). Table 1 presents the results obtained by our binary classifiers. It shows the overall classification accuracy as well as the F-scores for each class. From the table, it can be observed that the SVM classifier outperforms the simple MNB classifier in all three categories. In particular, the SVM classifier shows a very significant improvement for the ADR class F-score (an improvement of over 10%).

Table 1. Binary **c**lassification performances for the two classifiers: MNB and SVM.

Classifier	Accuracy (%)	ADR F-score	noADR F-score
MNB	77.6	0.540	0.852
SVM	82.6	0.652	0.884

4.2 Combining Classification Probabilities

Our final goal is to classify drugs to the normal or blackbox categories. We hypothesize that drugs in blackbox category exhibit more ADRs than normal drugs. Based on this assumption, we formulate our inference step as a probabilistic model in which we compute the probability of each drug belonging to one of the two categories, given the probabilities assigned to the comments by our automatic classifiers. Thus, for each comment, the only feature is the ADR/noADR probabilities assigned to the comment by the machine learning classifiers in the previous step. Each instance consists of all the probability estimates for a drug, and the target classes are the categories for the drugs (*i.e.*, blackbox or normal). Our inference step is given by the following equation.

$$\boldsymbol{P(y = N|x_1x_2 \ldots. x_n) = \frac{\sum_i^n P(y = noADR \mid x_i)}{n}}$$

$$\boldsymbol{P(y = B|x_1x_2 \ldots. x_n) = \frac{\sum_i^n P(y = \ ADR \mid x_i)}{n}}$$

Where, **N** stands for the Normal class and **B** stands for Blackbox class, $\boldsymbol{x_1x_2 \ldots. x_n}$ are the comments belonging to the drug being tested, and **n** is the number of comments for a drug. As mentioned, we derive the probability values for the above equation from the experiments described in Section 4.1.

This probability score models a uniform loss, in a sense that it assumes, the loss is equivalent if a normal drug is classified as blackbox or a blackbox drug is classified in normal class. The model favors the normal class as ADR comments are generally less in number compared to comments belonging to the noADR class. Therefore, when computing the sum of the two sets of probabilities for each drug, there are generally more noADR comments associated with the drugs, resulting in higher sums for noADR compared to ADR. This is a problem with the equation above as ADR comments should be held more decisive for the final classification than noADR comments. In order to incorporate the inherent bias we scaled ADR probability by a scaling factor α which is the ratio of the number of noADR comments to the number of ADR comments.

$$\boldsymbol{\alpha = \frac{\#\ of\ noADR\ comments}{\#\ of\ ADR\ comments}}$$

Thus, the final blackbox probability score is given as:

$$\boldsymbol{P(y = B|x_1x_2 \ldots. x_n) = \alpha * \frac{\sum_i^n P(y = \ ADR \mid x_i)}{n}}$$

Using this approach in the second step, if for a drug $P(y = B|x_1x_2 \ldots. x_n) * \ \alpha \quad > P(y = N|x_1x_2 \ldots. x_n)$, we categorize the drug as blackbox; otherwise, we categorize it as normal.

Table 2 presents the results of the second step of our two-step model. We use all comments associated with 20 normal and 18 blackbox drugs. For the results shown in the table, we use the classification probabilities of the SVM classifier from the first step. From the table it can be seen that the macro average F-score for our approach is 0.6. The recall and precision values are similar for both classes of drugs.

Table 2. Combining probability results using SVM probabilities for comments**.**

Average Precision	Average Recall	Macro Average F-score
0.611	0.61	0.60
Normal Precision	**Normal Recall**	**Normal F-score**
0.50	0.67	0.57
Blackbox Precision	**Blackbox Recall**	**Blackbox F-score**
0.72	0.56	0.63

We are interested in assessing if the classification accuracies/F-scores in the first step of our approach has an influence in the performance of the system in the second step. We hypothesize that if the classification performance of the first step can be improved, the overall performance of our approach can be improved as well. To investigate, we compare the results of the SVM probabilities with the MNB probabilities generated in the first step.

Table 3 shows the results of the drug classification approach when the MNB probability estimates are used. The average F-score in this case is 0.53, which is 7 points lower than the F-score when using the SVM classification probabilities. This suggests that the ADR/noADR classification accuracies of the first step do have an influence on the performance of the second step. In the first step (as shown in Table 1), the ADR F-score was over 10 points higher for the SVM classifier, and for the second step, the classification F-measure is 7 points higher. These results suggest that improvements in the ADR/noADR classification approach are likely to improve the detection of potentially harmful drugs. Moreover, this also suggests that the two-step approach that we propose is promising. Figure 2 presents two strip charts for each of the two classifiers showing the average ADR probabilities for the two sets of drugs. The figure shows that as classification accuracy increases in the first step, the separation between the two categories of drugs based on average ADR probabilities tends to get better (*i.e.*, blackbox drugs tend to have higher probabilities, on average, than normal drugs). However, at this point of research, and with the current annotated set we have, this separation is only marginal.

Table 3. Combining probability results using MNB probabilities for comments.

Average Precision	Average Recall	Macro Average F-score
0.59	0.68	0.53
Normal Precision	**Normal Recall**	**Normal F-score**
0.25	0.83	0.38
Blackbox Precision	**Blackbox Recall**	**Blackbox F-score**
0.94	0.53	0.68

Since our experiments so far indicate that the ADR *vs.* noADR classification accuracies play an important role in the automatic separation of blackbox and normal drugs, we are interested in investigating if the classification accuracies for that task can be improved if further data is annotated and made available for training. In particular, we are interested in analyzing how the ADR class F-score changes as the size of the training set is varied. We performed a number of experiments, each time reducing the size of the data set by 10% and applying the same 10-fold cross validation approach. Figure 3 shows how the classification performance changes as the size of the data set is increased. From the figure it can be seen that as the size of the training set increases, the ADR F-score shows sturdy improvement as well. The ADR F-score maintains a steady positive gradient till the far right of the graph, meaning that the availability of more training data should improve accuracy further. Thus, it is likely that as more annotated data is made available, the overall classification of drugs into final categories such as normal and blackbox can be improved.

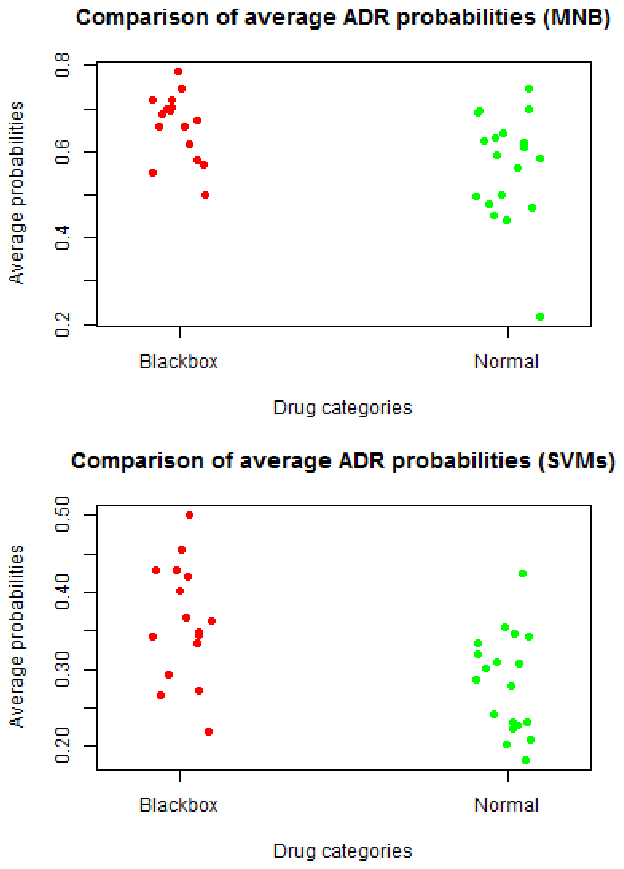

Figure 2. Strip charts showing how the average ADR probabilities for the blackbox and normal drugs are distributed for the MNB and SVM classifiers.

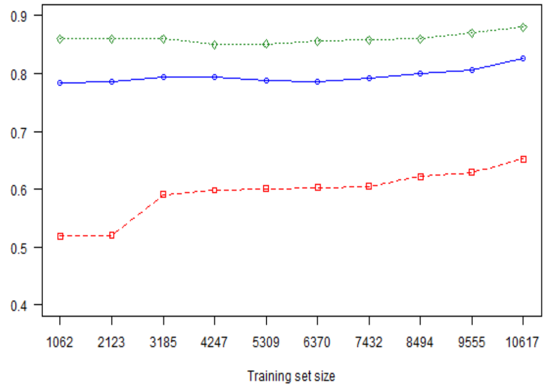

Figure 3. Classification performance *vs.* size of training set. The red line (bottom) indicates the ADR F-score, the green line (top) indicates the noADR F-score, and the blue line (middle) represents the overall accuracy.

Table 4. Top three drugs using multinomial SVM comment probabilities.

Normal (false positive)		Blackbox (false negative)	
Drug Name	**Score**	**Drug Name**	**Score**
Lyrica®	1.04275	Levaquin®	0.40738
Nicotrolinhaler®	0.99218	Baclofen	0.52334
Zetia®	0.87086	Avelox®	0.53998

5 DISCUSSION

We used 38 drugs for testing, 20 were normal and 18 had received a blackbox warning from the FDA. Table 4 shows the three drugs in the normal category with the highest average ADR scores (false positives) and three drugs in the blackbox category with the lowest average ADR scores (false negatives), as ranked by our scoring function using the SVM and MNB probability values. Both the classifiers give similar results for the ranking. Note that while for the blackbox groups these drugs indicate false negatives, the top three drugs in the normal category could be considered false positives. We now provide a brief discussion of our analysis on the key reasons behind the scores obtained by these drugs.

Nicotrolinhlaer® is a prescription nicotine replacement inhaler indicated in smoking cessation assistance. The comments for this drug are generally negative, for example "*didnt work at all , just made me nauseous*", "*this thing was just disgusting !*", "*Once when I was in hospital my dr told hubby to get for me , seeings how it was my only option I used it , but when I was released I was smoking before we made it out of the parking lot !*". As shown in the above comments, users mention either an ADR or a negative opinion about the drug. Our manual inspection revealed that negative comments are quite common for this product, and as a consequence the drug gets misclassified as a blackbox drug to our model.

Similarly, *Lyrica®* is a drug administered to treat pain caused by nerve damage due to diabetes.[10] Lyrica has a large number of comments in the corpus. For example: *"started taking on 10/2/10 So far feeling dizzy n lightheaded hoping it passes in time but don't like what I've read about it, giving it a trial run before I tell my Dr I can't handle the side effects n function", "dizzy very bad..feel like I am waling on a boat. Very tired same as off the meds so unsure of the cause. Makes me stumble over my own feet when I take it".* There are positive comments about the drug as well: *"I think its helping", "It has made big difference with my pain". Lyrica*, similar to nicotine, has a relatively high prevalence of common adverse effects including 10-28% of users experience dizziness. However, these common adverse effects do not rise to the level of box warning. In our model, the scaling factor causes the ADR probability score to increase, falsely classifying it as a blackbox drug.

Two of the false negatives from the models are *Levaquin* and *baclofen*. *Levaquin* is indicated to treat infections caused by bacteria. Comments such as "*This is really more useful, for me, when I have a much milder infection*", indicate a helpful effect of *Levaquin.* Since there are only 10 comments of *Levaquin* in the corpus, the classifier becomes inclined towards the normal class despite the scaling factor. Another factor contributing to the false negative response of *Levaquin* may be related to individual commenting. The boxed warning for *Levaquin* is related to tendinitis and tendon rupture especially over the age of 60 years (Seeger *et al.*, 2006). The small number of comments, and low prevalence of disease may have contributed to *Levaquin's* false negative classification. *Baclofen* is indicated to treat muscle spasticity. For Baclofen there are many comments indicating helpfulness of drug for pain relief, (*e.g.*, "*notice that it does help alot*", "*It helps with tremors*"). Hence the comments generally do not indicate ADR resulting in the misclassification of the drug as normal.

Although normal drugs such as *Nicotrolinhaler®*, *Lyrica®*, and *Zetia®* were classified as blackbox, the high prevalence of negative comments may provide important ADR signals for drugs that do not currently have FDA boxed warnings. We found that the larger the number of comments associated with a drug, the more stable its prediction is. Thus, if larger numbers of comments were available for all the drugs, we would perhaps be able to make more accurate predictions.

6 CONCLUSION

In this paper, we proposed an approach for classifying drugs into normal and blackbox categories, based on the automatic classification of comments associated with them extracted from social media. This classification is based on our hypothesis that blackbox drugs show more ADRs than normal drugs. We applied a two-step approach, first to classify comments into ADR *vs*. noADR categories. We then utilized those classifications to categorize drugs into the two above-mentioned categories. The results obtained, while promising regarding the individual classification of comments as ADR or noADR, are marginal with respect to the overall classification of the drug: distinguishing true signals from noise when utilizing consumer-generated comments from social media for post-marketing surveillance.

[10] http://www.webmd.com/drugs/drug-93965-Lyrica+Oral.aspx?drugid=93965&drugname=Lyrica+Oral

However, given the novelty of the idea, the approach holds promise, particularly as more training data is made available.

REFERENCES

Aramaki, E., Miura, Y., & Tonoike, M. (2010). Extraction of Adverse Drug Effects from Clinical Records. *Studies in Health*, 739–743. doi:10.3233/978-1-60750-588-4-739

Benton, A., Ungar, L., Shawndra, H., Hennessy, S., Mao, J., Chung, A., … Holmes, J. H. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics*, 989–996.

Carletta, J. (1996). Squibs and Discussions Assessing Agreement on Classification Tasks : The Kappa Statistic. *Computational Linguistics*.

Chee, B. W., Berlin, R., & Schatz, B. (2011). Predicting adverse drug events from personal health messages. *AMIA Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2011*, 217–26.

Friedman, C. (2009). Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of the Electronic Health Record. *AIME '09 Proceedings of the 12th Conference on Artificial Intelligence in Medicine: Artificial Intelligence in Medicine, 2009*, 1–5.

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., … Gonzalez, G. (2014). Mining Twitter for Adverse Drug Reaction Mentions : A Corpus and Classification Benchmark. In *LREC, BioTxtM 2014*.

Guerini, M., Gatti, L., & Turchi, M. (2013). Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. *arXiv Preprint arXiv:1309.5843 (2013).*

Gurulingappa, H., Rajput, A., & Toldo, L. (2012). Extraction of Adverse Drug Effects from Medical Case Reports. *Drugs*, 1–4.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, *6*, 343. doi:10.1038/msb.2009.98

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., & Gonzalez, G. (2010). Towards Internet-Age Pharmacovigilance : Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2010.*, (July), 117–125.

Nikfarjam, A., & Gonzalez, G. H. (2011). Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2011*, 1019–26. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243273&tool=pmcentrez&rendertype=abstract

Niu, Y., Zhu, X., Li, J., & Hirst, G. (2005). Analysis of Polarity Information in Medical Text University of Toronto. *AMIA Annual Symposium Proceedings. Vol. 2005. American Medical Informatics Association, 2005.*, *2005*(August 2001), 570–574. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16779104

Harpaz, R., DuMouchel, W., Shah, N., Madigan, D., Ryan, P., and Friedman, C., (2012). Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics 91.6*, 1010–1021. Retrieved from http://www.nature.com.ezproxy1.lib.asu.edu/clpt/journal/v91/n6/abs/clpt201250a.html

Sampathkumar, H., Luo, B., & Chen, X. (2012). Mining Adverse Drug Side-Effects from Online Medical Forums. *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, *3*(1), 150–150. doi:10.1109/HISB.2012.75

Sarker, A., Molla, D., & Paris, C. (2013). Automatic Prediction of Evidence-based Recommendations via Sentence-level Polarity Classification. *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Nagoya, Japan.*, (October), 712–718. Retrieved from http://aclweb.org/anthology/I/I13/I13-1084.pdf

Seeger, J. D., West, W. a, Fife, D., Noel, G. J., Johnson, L. N., & Walker, A. M. (2006). Achilles tendon rupture and its association with fluoroquinolone antibiotics and other potential risk factors in a managed care population. *Pharmacoepidemiology and Drug Safety*, *15*(11), 784–92. doi:10.1002/pds.1214

Toldo, L., Bhattacharya, S., & Gurulingappa, H. (2012). Automated identification of adverse events from case reports using machine learning. *Proceedings XXIV Conference of the European Federation for Medical Informatics. Workshop on Computational Methods in Pharmacovigilance, Pisa, Italy. 2012.*

Wang, X., Hripcsak, G., Markatou, M., & Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association : JAMIA*, *16*(3), 328–37. doi:10.1197/jamia.M3028

World Health Organization. (2013). The Importance of Pharmacovigilance: Safety Monitoring of Medicinal Products. 2002. *World Health Organization; Geneva*.

Yates, A., & Goharian, N. (2013). ADRTrace : Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. *Advances in Information Retrieval. Springer Berlin Heidelberg, 2013. 816-819.*, 816–819.

Yeganova, L., Comeau, D. C., Kim, W., & Wilbur, W. J. (2011). Text Mining Techniques for Leveraging Positively Labeled Data. *Proceedings of BioNLP 2011 Workshop. Association for Computational Linguistics, 2011.*, (Zhang 2004), 155–163.

Zeng-Treitler, Q., Goryachev, S., Tse, T., Keselman, A., & Boxwala, A. (2008). Estimating consumer familiarity with health terminology: a context-based approach. *Journal of the American Medical Informatics Association : JAMIA*, *15*(3), 349–56. doi:10.1197/jamia.M2592