Automated Detection of Adverse Drug Reactions From Social Media Posts With Machine Learning

Ilseyar Alimova and Elena Tutubalina

Laboratory of Chemoinformatics and Molecular Modeling
Kazan (Volga Region) Federal University, Kazan, Russia
alimovailseyar@gmail.com, elvtutubalina@kpfu.ru

Abstract. Adverse drug reactions can have serious consequences for patients. Social media is a source of information useful for detecting previously unknown side effects from a drug since users publish valuable information about various aspects of their lives, including health care. Therefore, detection of adverse drug reactions from social media becomes one of the actual tools for pharmacovigilance. In this paper, we focus on identification of adverse drug reactions from user reviews and formulate this problem as a binary classification task. We developed a machine learning classifier with a set of features for resolving this problem. Our feature-rich classifier achieves significant improvements on a benchmark dataset over baseline approaches and convolutional neural networks.

Keywords: adverse drug reactions, text mining, health social media analytics, machine learning, deep learning

1 Introduction

Detection of drug side effects is one of the main tasks in the pharmacy industry. Before the release of the medication, a number of clinical trials are conducted in order to detect side effects, which further fit into the drug's instructions. However, clinical trials do not allow to identify all drug side effects, because some of them appear after long-term use of the drug or have an effect only on a certain group of patients who did not participate in clinical trials. Recent work have shown that 462 medical products were withdrawn from sales between 1950 and 2014 years [1]. Moreover, side effects appeared in the post-approval period can cause serious problems to human health and even lead to death. [2–5]. The detection of drug side effects in post-approval periods is a difficult challenge for pharmacovigilance.

One of the methods of finding new side effects for released into the market drugs is social media analysis [6]. With the development of social networks, users often write about the problems associated with taking medications on Twitter and various forums related to health and drugs. Manual processing such a volume of text information is impossible, therefore methods of natural language

processing are widely used for automatic text processing [7–10] including to extract information about adverse effects, as evidenced by a number of review articles on this topic [11–15].

One of the tasks of detection of Adverse Drug Reactions (ADR) is to classify any disease-related information. This classification task is necessary to remove noise information and detect if the text contains mentions of side effects. Wee formulatd the problem as a binary classification to determine whether this side effect is adverse or not. The first class includes all the drug effects that had a negative impact on a health. The other group includes indications, disease symptoms, beneficial effects and effects experienced not by the patient directly. We developed a machine learning classifier with a set of features for resolving this problem. We used word embedding features, including vector representation and brown clusters trained on social media posts on the topic of health and drugs. We will demonstrate the effectiveness of our approach on message-level and entity-level classification. For message-level classification, we tested our approach on a benchmark dataset of tweets named the Twitter corpus [16]. For entity-level, we tested our approach on a benchmark dataset of user reviews named the CSIRO Adverse Drug Event Corpus (CADEC) [17]. We used Logistic Regression and Linear Support Vector Machine (SVM) classifiers. Our feature-rich classifiers outperformed Convolutional Neural Networks (CNN) designed for text classification [18] and a state-of-the-art approach based on SVM introduced by Sarker et al. [16].

The rest of the paper is structured as follows. In Section 2, we discuss related work. In Section 3, we describe our classifier with hand-crafted features. Section 4 provides evaluation results. Section 5 concludes this paper.

2 Related work

Many previous works have been devoted to the detection of ADR. In Social Media Mining Shared Task 2016 [19], the task 1 was to classify user posts whether a user discusses ADRs. Participants were given a marked body of tweets on the topic of health, their task was to determine whether there was information about side effects in the text of the tweet or not. The winner system [20] is based on Random Forest and used words, the co-occurrence of drug and side effect, sentiment, negation, change in tweets tone and question in tweets features and obtained ADR class F-score of 41.95%. The second place system used different configurations of Maximum Entropy Classifier and concept-matching classier based on ADR lexicon and obtained ADR class F-measure of 41.82% [21]. Ofoghi et. al. [22] introduced SVM-based classifier with a sentiment, emotion classes, mention from Unified Medical Language System (UMLS), chemicals/drugs/diseases lexicon based features and got ADR class F-measure of 35.8%. The fourth-placed system applied SVM classifier with syntactic, lexicon, polarity and topic modeling based features and performed ADR class F-measure of 33% [23]. The last system from top five also developed the SVM-based system with n-Gram, word embedding, cluster, lexical features and obtained F-measure

of 31.74% [24]. SVM is the most popular text classification techniques, which has been widely adopted in patient social media research [25–28]. Sarker et. al. [16] tried Maximum Entropy model with n-gram, UMLS semantic types and concept IDs, syn-set expansion, change phrases, ADR lexicon matches, SentiWordNet scores, topic-based features and obtained ADR class F-measures of 81.2%, 53.8% and 67.8% for Twitter[1], DailyStrength [28] and ADR corpora [29], respectively. Liu et. al [30] developed a rule-based algorithm to identify adverse drug events and event pairs from all related drugs. The classifier obtained macro-averaged F-measure of 69.20%. Trung et. al. [31] introduced CNN for classification of ADRs. CNN with pre-trained word embeddings showed better results over machine learning classifiers with macro-averaged F-measures of 51% and 87% on Twitter [16] and the second corpus of adverse events [32], respectively.

As you can see, basically, for the task ADR classification, machine-learning approaches were used with similar sets of features, including n-grams, ADR lexicon, sentiment features, a presence of drugs, UMLS mentions. Only in [24] were used embedding clusters features, trained with word2vec on unlabeled user reviews about drugs, collected from Twitter and Daily Strength, and k-means Clustering from [33].

3 The Proposed Method

We applied two machine learning models based on Linear SVM and Logistic Regression with entity-level and context-level sets of features. Entity-level features were applied only for entity tokens. Context-level features were used within a window of four words on each side of an entity, including the entity itself. For tweets from the Twitter corpus, we applied the features described below to all tokens in the tweets text. The methods were implemented with classes from the scikit-learn library [34]: LinearSVC and Logistic Regression (with parameters class_weight='auto' and penalty='l2'). The source code of our classifier can be found here[2].

3.1 Features

Context-level features:

- Bag of words (bow): we used unigrams and bigrams.
- Part of Speech (PoS): we counted the number of nouns, verbs, adverbs and adjectives.
- Sentiment features (sent): we applied state-of-the-art lexicon features for sentiment analysis described in [35]. we used SentiWordNet [36], MPQA Subjectivity Lexicon [37], Bing Liu's dictionaries [38].
- Pointwise mutual information (pmi): we counted PMI for the large corpus of user reviews named the Health corpus collected from various resources and described further. We used the PMI score as a feature.

[1] http://diego.asu.edu/index.php?downloads=yes
[2] https://github.com/Ilseyar/adr_classification

- Drugname and ADR presenting (drug_adr): this is a binary vector of length two. The first component of the vector shows the presence of the name of the drug from FDA[3], the second is the presence of ADR effect from the dictionary COSTART[4].
- Emoticons (emot): a binary vector of length 2, showing the presence of positive and negative emoticions.

Entity-level features:

- Word embedding (emb): we used vector representation trained on social media posts from [39]. We calculated the average of the vectors of dimension 200 for each token and applied it as a feature.
- Cluster-based representation (cls): we used clusters computed in [39] with Brown hierarchical clustering algorithm and represented each entity as a binary vector of dimension 150.
- Semantic Types from Unified Medical Language System (umls): we used UMLS version 2.0 and counted the number of tokens from each UMLS semantic types. UMLS semantic types are subject categories that provide a categorization of all concepts represented in the UMLS. For example, clinical drug, medical device, vitamin etc.

Word embedding vectors were obtained with using word2vec trained on unlabeled Health corpus consists of 1,180,080 reviews from askapatient.com[5], dailystrength.org[6], drugscom.com[7], amazon health-related dataset[8], webmd.com[9]. The parameters of word embeddings are vectors with size of 200, the length of the local context of 10, the negative sampling of 5, vocabulary cutoff of 10, Continuous Bag of Words model [39].

4 Experiments

In this section, we describe our experiments with feature-rich classifiers and deep learning models.

4.1 Evaluation Datasets

We conducted experiments on CADEC [17] and Twitter corpora [16].

CADEC The CADEC corpus consists of annotated user reviews from the askapatient.com medical forum. There are five types of annotations: drug, adverse effect, disease, symptom, and finding. The 'drug' label was given to all drug

[3] https://www.fda.gov/
[4] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/
[5] http://www.askapatient.com/
[6] https://www.dailystrength.org/
[7] https://www.drugs.com/
[8] http://jmcauley.ucsd.edu/data/amazon/
[9] http://www.webmd.com/

names in the text. All side effects that are directly related to the drug, which was written about in the review, are annotated as 'ADR'. The entity 'disease' is the indication to taking the drug. The entity labeled as 'symptom' specifies the indication disease. The label 'finding' was given to any side effects or symptoms, which are not related to the patient, and for entities that annotators could not establish belonging to one of the classes. We grouped diseases, symptoms and findings as single class called 'Other'. The corpus contains 6320 entities, 5770 of them marked as 'ADR'.

Twitter The Twitter corpus contains user tweets on the topic of health and adopted from [16]. Each tweet labeled whether tweets text contains the information about adverse drug reactions. Since the policy of Twitter does not allow the publication of tweet texts in the public domain, the corpus consists of a file containing the id tweet, the user id, and the class number. The creators of the corpus published a script for downloading tweet texts. A fraction of the tweets (36%) was no longer available on Twitter, which made our results are not directly comparable to the ones of previous works. During pre-processing, we removed all URLs, user mentions and symbols of re-tweets using the tweet-preprocessor package[10].

4.2 Baseline Methods

We compare our approach with two baselines:

- **SVM from [16]**: The developed method based on SVM with Linear kernel. Features applied in this method incorporated 1,2,3-grams, synsets, sentiment, change phrases, ADR lexicon, topic-based features, the lengths of the text segments in words, the presence of comparatives and superlatives adjectives and modal verbs. The synsets features consist of synonyms for each adjective, noun or verb in a sentence obtained from WordNet. For sentiment feature, the following dictionaries were used: SentiWordNet [36], MPQA Subjectivity Lexicon [37], Bing Liu's dictionary [38]. The change phrases feature presented a vector of dimension 4, where each component shows the number of words belonging to 'less', 'more', 'good', 'bad' words dictionaries. ADR lexicon feature used SIDER[11], Consumer Health Vocabulary[12], COSTART[13] and DIEGO_LAB[14] dictionaries and consists of two parameters, the first shows the token belonging to the ADR dictionary, the second number of tokens from the ADR dictionary. The topic-based feature consists of the topic terms that appear in the instance and the sums of all the relevance scores of the terms in each instance. We used the publicly available code of this method[15].

[10] https://pypi.python.org/pypi/tweet-preprocessor/0.4.0
[11] http://sideeffects.embl.de/
[12] http://www.consumerhealthvocab.org/
[13] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/
[14] http://diego.asu.edu/Publications/ADRClassify.html
[15] https://bitbucket.org/asarker/adrbinaryclassifier/downloads/

– **CNN**: in order to get local features from a review with CNNs we have used multiple filters of different lengths [18]. Pooled features are fed to a fully connected feed-forward neural network (with dimension 100) to make an inference, using rectified linear units as output activation. Then we apply a softmax classifier with a number of outputs equals 2. We applied dropout rate of 0.5 [40] to the fully connected layer. We trained CNN for 10 epochs since CNN achieved lower results after ten epochs. Embedding layers are trainable for all networks; this setting leads to a significant gain in performance. We set mini-batch size to 128 with the Adam optimizer [41]. We found 97% and 84% of words in the vocabulary in the CADEC corpus and the Tweet corpus, respectively, and for other words, the representations were uniformly sampled from the range of embedding weights [42]. We used the Keras library[43] for implementation. Both CNN and our classifier used the same word embeddings trained on health-related comments from [39].

4.3 Results and Analysis

We performed pre-processing by lower-casing all words. We tested the methods on the 5-folds cross validation. We computed macro-averaged recall (R), precision (P) and F_1-measures (F). Table 1 presents the variances of the F1-measure in our cross-validation results. We took the best configuration of features for each model, bow, pos, sent, cls, umls for Linear SVM and all set of described features for Logistic Regression model.

Table 1. 5-fold cross-validation performances of our feature-rich methods and baselines.

Method	Corpus	Folds (F_1-measure)					av. F
		1	2	3	4	5	
SVM + best config.	CADEC	.783	**.788**	**.817**	.783	**.846**	**.803**
Logistic Regression + best config.	CADEC	**.787**	.767	.791	.785	.782	.783
CNN, [1, 2] filters	CADEC	.760	.771	.792	**.805**	.784	.782
CNN, [1, 2, 3] filters	CADEC	.740	.760	.770	.797	.777	.769
CNN, [2, 3, 4] filters	CADEC	.771	.739	.799	.787	.757	.771
CNN, [1, 2, 3, 4] filters	CADEC	.765	.766	.773	.804	.785	.779
CNN, [1, 2, 3, 4, 5] filters	CADEC	.796	.773	.768	.794	.787	.783
Classifier from [16]	CADEC	.645	.676	.737	.677	.703	.688
SVM + best config.	Twitter	.683	.715	.706	.700	.709	.702
Logistic Regression + best config.	Twitter	**.711**	**.752**	.747	**.737**	**.738**	**.737**
CNN, [1, 2] filters	Twitter	.628	.735	.739	.714	.695	.702
CNN, [1, 2, 3] filters	Twitter	.635	.716	.726	.708	.725	.702
CNN, [2, 3, 4] filters	Twitter	.692	.703	.722	.712	.651	.696
CNN, [1, 2, 3, 4] filters	Twitter	.695	.705	**.749**	.674	.642	.693
CNN, [1, 2, 3, 4, 5] filters	Twitter	.677	.701	.725	.694	.693	.698
Classifier from [16]	Twitter	.676	.677	.691	.680	.698	.684

Table 2. Features impact evaluation on the CADEC and Twitter corpora with Linear SVM and Logistic Regression models respectively with different groups of features.

	CADEC			Twitter		
Features	**P**	**R**	**F**	**P**	**R**	**F**
bow	.827	.740	.775	.642	**.751**	.666
bow, pos	.824	.743	.776	.722	.682	.700
bow, pos, sent	.823	.745	.777	.719	.698	.708
bow, pos, sent, cls	.832	.776	.788	.721	.708	.714
bow, pos, sent, cls, umls	**.844**	.773	**.803**	.721	.708	.714
bow, pos, sent, cls, umls, pmi	.839	.770	.799	.723	.710	.716
bow, pos, sent, cls, umls, pmi, emb	.822	**.777**	.797	**.730**	.742	.736
bow, pos, sent, cls, umls, pmi, emb, drug_adr	.842	.772	.802	.728	.745	.736
bow, pos, sent, cls, umls, pmi, emb, drug_adr, emot	.812	.774	.792	.729	.746	**.737**

The classification results (especially F1-measure) in Table 1 indicate the advantage of machine learning classifiers. The feature-rich SVM and Logistic Regression classifiers achieved best results on the CADEC corpus and the Twitter corpus, respectively. Therefore, classical machine learning approaches with rich additional information can still outperform neural network approaches for domain-specific problems like detection of adverse drug reactions.

We also investigated the effectiveness of features in Table 2. As can be seen from the tables, Linear SVM with features obtained the maximum value of the macro F-measure 80.3% on CADEC corpus and Logistic Regression got macro F-measure 73.7% on the Twitter corpus. For SVM, cluster-based features and umls increase significantly the effectiveness of the classifier on the CADEC corpus. For Logistic Regression, most significant feature is word embeddings.

We also investigated the effectiveness of different sentiment lexicons. The sentiment feature was computed only for the Bing Liu's lexicon since the full set of sentiment lexicons didn't improve the performance. We tried to extend our set of features with ADR lexicons similar to [16]. We used COSTART[16], SIDER[17] and DIEGO Lab ADR Lexicon from [16] with the best configurations of features. The results of these experiments are presented in Table 3. According to these, only COSTART lexicon improved classification results for Twitter corpus.The possible explanation is that the dictionaries provide poor coverage of the corpora, as shown in the Table 4

4.4 Error analysis

In this section, we present an analysis of classification errors. We looked at 150 examples of each type of errors and identified the main causes of errors with the examples presented in the Table 5. The other cases are difficult to combine into large groups, each of them requires more detailed consideration.

[16] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/
[17] http://sideeffects.embl.de/

Table 3. Lexicon features impact evaluation on the CADEC and Twitter corpora with Linear SVM and Logistic Regression models respectively with groups of features with the best results.

	CADEC			**Twitter**		
Lexicons	**P**	**R**	**F**	**P**	**R**	**F**
COSTART	.842	.767	.799	.729	.744	.736
SIDER	.840	.768	.799	.729	.739	.734
ADR Lexicon	.844	.769	.801	.719	.698	.708

Table 4. Summary of statistics of considered lexicons

Corpus	unigrams	COSTART	SIDER	ADR lexicon
CADEC	3204	111	145	311
Twitter	11929	95	149	315

CADEC corpus

- *ADR with pain word.* Most errors are associated with the entities that include word 'pain' (30%). If the word 'pain' was next to the words denoting negative sentiment, the system erroneously classified it as 'Adverse', however, it becomes clear from the context that in this case, this is not an adverse drug reaction. On the contrary, if the entity 'pain' was encountered in a positive context, the system classified these cases as 'Other', but often in such cases, the patient described his condition after he stopped taking the medicine and the annotators labeled it as 'Other'.
- *Training set disbalance.* Some errors appeared because of the disbalance in the training set. For example, the entities: depression, swelling, headache, cramps most often belong to the class 'Adverse', and the words inflamation, fibromyalgia, MS are mostly classified as 'Other'. The error associated with this is more common for the case where the system incorrectly labeled entities as Adverse (about 18%) and for another type of errors it is only 5%.
- *ADR context.* The presence of terms denoting the adverse drug reaction next to the entity, caused the system to erroneously give the answer 'Adverse' instead of right 'Other'. It caused 25% of errors.

Twitter corpus

- *No ADR mention.* The absence of mention of the concrete adverse effect is the most common reason of the erroneous decision of our system to label the tweet as 'Other'. This is 40% of all cases. Sometimes such cases describe a patient's violation of a diet, in particular, alcohol consumption during the drug taking, or overdose.
- *Adverse drug context.* The second most common error (about 40%) is associated with the mention of the drug and the side effects together in the text

of the tweet. While the system determined it to the 'Adverse' class, the right answer was 'Other' because the effects did not apply to the drug.

- *No ADR description.* In 40% of erroneously classified tweets, users wrote about the presence of ADR but did not describe them specifically. In this case, the system classified this tweet as 'Adverse', however, in the gold file such cases was annotated as 'Other'.
- *Not own user experience.* In 15% of cases, the user described the side effect of the drug that he did not experience, in this cases our system also recognized this tweet as 'Adverse', although, the correct answer was 'Other'.
- *Positive side effect.* If the adverse effect was identified as positive by user the system defined the tweet in the category 'Other' class, however in the gold file it was labeled as 'Adverse' (10%).
- *No drug name.* The lack of mention of the drug name also led to an error when the system incorrectly classified a tweet to 'Other' (10%).

Table 5. Examples illustrating common reasons behind the misclassification

Review's text	Corpus	Classified as	Issues
If I miss a day, headaches begin to creep in.	CADEC	Adverse	Training set disbalance
...depression worse, fibromyalgia much worse.	CADEC	Other	Training set disbalance
your tweets are depressing me. haha. paxil	Twitter	Adverse	adverse drug context
...by an adverse reaction to the drug effexor.	Twitter	Adverse	no ADR description
...12 rivaroxaban diary: headache, right shoulder...	Twitter	Adverse	Not own user experience
i've had no appetite since i started on prozac, i guess that's a good thing	Twitter	Adverse	Positive side effect
nicotine lozenges. if i go cold turkey i cant think (or see) straight ...	Twitter	Adverse	No drug name
...feel like this fluoxetine is messing with my perspective time	Twitter	Other	No ADR mention

5 Conclusion

In this paper, we have focused on the ADR classification task. We have explored Linear SVM and Logistic Regression classifiers with a rich set of features including sentiment and semantic features, word embeddings, and lexicon features. We tested the proposed approach on two benchmark corpora of user reviews and tweets and compared with the state-of-the-art classifier and convolutional neural networks.

We demonstrated the superiority of machine learning approach as compared to a convolutional neural network and another previously proposed approach. The most improvement for ADR classification gave sentiment and word embedding features. We also showed that the features based on ADR lexicon do not

give significantly improve for classification results. The possible explanation is that a dictionary of ADRs is specific for a particular group of drugs (e.g., weight gain or loss is a side effect of depression treatment and the reason for taking orexigenic drugs and appetite suppressants). Hence, in further work, we plan to create drug-specific dictionaries and incorporate them into neural models.

Acknowledgments

Work on problem definition and neural networks was carried out by Elena Tutubalina and supported by the Russian Science Foundation grant no. 15-11-10019. Other parts of this work were performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. Onakpoya, I.J., Heneghan, C.J., Aronson, J.K.: Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. BMC medicine **14**(1) (2016) 10
2. Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A.K., Walley, T.J., Farrar, K., Park, B.K., Breckenridge, A.M.: Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. Bmj **329**(7456) (2004) 15–19
3. Classen, D.C., Pestotnik, S.L., Evans, R.S., Lloyd, J.F., Burke, J.P.: Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. Jama **277**(4) (1997) 301–306
4. Lazarou, J., Pomeranz, B.H., Corey, P.N.: Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama **279**(15) (1998) 1200–1205
5. Bates, D.W., Cullen, D.J., Laird, N., Petersen, L.A., Small, S.D., Servi, D., Laffel, G., Sweitzer, B.J., Shea, B.F., Hallisey, R., et al.: Incidence of adverse drug events and potential adverse drug events: implications for prevention. Jama **274**(1) (1995) 29–34
6. Sloane, R., Osanlou, O., Lewis, D., Bollegala, D., Maskell, S., Pirmohamed, M.: Social media and pharmacovigilance: a review of the opportunities and challenges. British journal of clinical pharmacology **80**(4) (2015) 910–920
7. Tutubalina, E., Nikolenko, S.: Automated prediction of demographic information from medical user reviews. In: International Conference on Mining Intelligence and Knowledge Exploration, Springer, Cham (2016) 174–184
8. Solovyev, V., Ivanov, V.: Knowledge-driven event extraction in russian: corpus-based linguistic resources. Computational intelligence and neuroscience **2016** (2016) 16
9. Sayfullina, L., Eirola, E., Komashinsky, D., Palumbo, P., Karhunen, J.: Android malware detection: Building useful representations. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). (Dec 2016) 201–206
10. Ivanov, V., Tutubalina, E., Mingazov, N., Alimova, I.: Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars. In: Proceedings of International Conference Dialog. Volume 2. (2015) 22–34

11. Murff, H.J., Patel, V.L., Hripcsak, G., Bates, D.W.: Detecting adverse events for patient safety research: a review of current methodologies. Journal of biomedical informatics **36**(1) (2003) 131–143
12. Sarker, A., Ginn, R., Nikfarjam, A., OConnor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G.: Utilizing social media data for pharmacovigilance: A review. Journal of biomedical informatics **54** (2015) 202–212
13. Lardon, J., Abdellaoui, R., Bellet, F., Asfari, H., Souvignet, J., Texier, N., Jaulent, M.C., Beyens, M.N., Burgun, A., Bousquet, C.: Adverse drug reaction identification and extraction in social media: a scoping review. Journal of medical Internet research **17**(7) (2015) e171
14. Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., Shah, N.H.: Text mining for adverse drug events: the promise, challenges, and state of the art. Drug safety **37**(10) (2014) 777–790
15. Harpaz, R., DuMouchel, W., Shah, N.H., Madigan, D., Ryan, P., Friedman, C.: Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics **91**(6) (2012) 1010–1021
16. Sarker, A., Gonzalez, G.: Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics **53** (2015) 196–207
17. Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C.: Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics **55** (2015) 73–81
18. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
19. Sarker, A., Nikfarjam, A., Gonzalez, G.: Social media mining shared task workshop. In: Proceedings of the Pacific Symposium on Biocomputing. (2016) 581–592
20. Rastegar-Mojarad, M., Komandur Elayavilli, R., Yu, Y., Hiu, H.: Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing. (2016)
21. Zhang, Z., Nie, J., Zhang, X.: An ensemble method for binary classification of adverse drug reactions from social media. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing. (2016)
22. Ofoghi, B., Siddiqui, S., Verspoor, K.: Read-biomed-ss: Adverse drug reaction classification of microblogs using emotional and conceptual enrichment. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing. (2016)
23. Jonnagaddala, J., Jue, T.R., Dai, H.: Binary classification of twitter posts for adverse drug reactions. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing. (2016) 4–8
24. Egger, D., Uzdilli, F., Cieliebak, M., Derczynski, L.: Adverse drug reaction detection using an adapted sentiment classifier. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing. (2016)
25. Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., OConnor, K., Sarker, A., Smith, K., Gonzalez, G.: Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In: Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing, Citeseer (2014)
26. Yang, M., Wang, X., Kiang, M.Y.: Identification of consumer adverse drug reaction messages on social media. In: PACIS. (2013) 193

27. Bian, J., Topaloglu, U., Yu, F.: Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 international workshop on Smart health and wellbeing, ACM (2012) 25–32
28. Patki, A., Sarker, A., Pimpalkhute, P., Nikfarjam, A., Ginn, R., OConnor, K., Smith, K., Gonzalez, G.: Mining adverse drug reaction signals from social media: going beyond extraction. Proceedings of BioLinkSig **2014** (2014) 1–8
29. Gurulingappa, H., Mateen-Rajpu, A., Toldo, L.: Extraction of potential adverse drug events from medical case reports. Journal of biomedical semantics **3**(1) (2012) 15
30. Liu, X., Liu, J., Chen, H.: Identifying adverse drug events from health social media: a case study on heart disease discussion forums. In: International Conference on Smart Health, Springer (2014) 25–36
31. Huynh, T., He, Y., Willis, A., Rüger, S.: Adverse drug reaction classification with deep neural networks, COLING (2016)
32. Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L.: Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of biomedical informatics **45**(5) (2012) 885–892
33. Nikfarjam, A., Sarker, A., OConnor, K., Ginn, R., Gonzalez, G.: Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association **22**(3) (2015) 671–681
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(Oct) (2011) 2825–2830
35. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research **50** (2014) 723–762
36. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: LREC. Volume 10. (2010) 2200–2204
37. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics (2005) 347–354
38. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2004) 168–177
39. Miftahutdinov, Z., Tutubalina, E., Tropsha, A.: Identifying disease-related expressions in reviews using conditional random fields. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii **1**(16) (2017) 155–166
40. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1) (2014) 1929–1958
41. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
42. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. (2015) 1026–1034
43. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)