SOCIAL MEDIA MINING SHARED TASK WORKSHOP

ABEED SARKER*
Department of Biomedical Informatics, Arizona State University,
Scottsdale, AZ 85259, United States of America
**E-mail: abeed.sarker@asu.edu*
diego.asu.edu

AZADEH NIKFARJAM
Department of Biomedical Informatics, Arizona State University,
Scottsdale, AZ 85259, United States of America
E-mail: anikfarj@asu.edu

GRACIELA GONZALEZ
Department of Biomedical Informatics, Arizona State University,
Scottsdale, AZ 85259, United States of America
E-mail:graciela.gonzalez@asu.edu

Social media has evolved into a crucial resource for obtaining large volumes of real-time information. The promise of social media has been realized by the public health domain, and recent research has addressed some important challenges in that domain by utilizing social media data. Tasks such as monitoring flu trends, viral disease outbreaks, medication abuse, and adverse drug reactions are some examples of studies where data from social media have been exploited. The focus of this workshop is to explore solutions to three important natural language processing challenges for domain-specific social media text: (i) text classification, (ii) information extraction, and (iii) concept normalization. To explore different approaches to solving these problems on social media data, we designed a shared task which was open to participants globally. We designed three tasks using our in-house annotated Twitter data on adverse drug reactions. Task 1 involved automatic classification of adverse drug reaction assertive user posts; Task 2 focused on extracting specific adverse drug reaction mentions from user posts; and Task 3, which was slightly ill-defined due to the complex nature of the problem, involved normalizing user mentions of adverse drug reactions to standardized concept IDs. A total of 11 teams participated, and a total of 24 (18 for Task 1, and 6 for Task 2) system runs were submitted. Following the evaluation of the systems, and an assessment of their innovation/novelty, we accepted 7 descriptive manuscripts for publication— 5 for Task 1 and 2 for Task 2. We provide descriptions of the tasks, data, and participating systems in this paper.

Keywords: Concept Extraction; Text Classification; Adverse Drug Reaction; Pharmacovigilance; Social Media Mining.

1. Background

Adverse drug reactions (ADRs), defined as accidental injuries resulting from correct medical drug use, present a serious and costly health problem contributing to 5.3% of all hospital admissions each year.[1] The process of detection, assessment, understanding, and prevention of these events is called pharmacovigilance.[2] To facilitate pharmacovigilance efforts, governments worldwide have diverse surveillance programs. One example, in the U.S., is MedWatch;[a]

[a] `http://www.fda.gov/Safety/MedWatch/default.htm` [Accessed Sep-28-2015]

it enables both patients and providers to manually submit ADR information. However, these programs are chronically underutilized. A systematic review encompassing 12 countries, estimated an 85-94% under-reporting rate of ADRs in local, regional, and national level reporting systems. To improve detection rates, researchers have begun turning to alternative sources of healthcare data, such as social media.[3,4] Recent studies suggest that 26% of adult Internet users discussed personal health issues online, with 42% of them discussing current conditions on social media and 30% reportedly changing their behavior as a result.[5] Studies have focused on automatic classification of ADR assertive user posts,[6–10] and the automatic extraction of ADR mentions from posts.[11–14] However, despite the proposal of various techniques for utilizing social media data in the past, public availability of data is scarce, making direct comparison of different approaches impossible. Our recent release of large annotated data sets prepared from Twitter data[8,10,15] has opened up the possibility to compare the performances of distinct approaches for social media based pharmacovigilance.

1.1. *Shared task and workshop design*

To facilitate research on social media based pharmacovigilance and social media text mining in general, we organized this workshop on social media text mining. Unlike traditional workshops, where manuscripts within the scope of the workshop are submitted, reviewed and chosen for acceptance, we ran this workshop as a shared task. The shared task consisted of three tasks: classification, extraction and normalization. We provided annotated data for the three tasks, and participants were required to develop and submit systems for evaluation on previously unreleased test data. Following the evaluation of the submitted systems, participants were required to submit short system descriptions. The system descriptions were reviewed by at least one peer and one member of the workshop organizing committee for selection.

In total, 11 teams registered for the shared task, and a total of 24 system runs were submitted. We received 18 submissions for Task 1, and 6 for Task 2. Unfortunately, perhaps due to the complex nature of the task, we did not receive any system submissions for Task 3. 9 teams were invited to submit system descriptions, of which 7 were eventually selected for publication.

In the following sections, we provide detailed descriptions of the task and submitted systems. For the rest of this section, we provide some background on social media, its use in the public health domain in general, and the challenges faced by text mining systems relying on this source of data.

1.2. *Pharmacovigilance from social media*

Over the last few years, social networks have seen massive growths (*e.g.*, as of 29th September 2015, Twitter has over 645,750,000 users and grows by an estimated 135,000 users every day, generating 58 million tweets per day[b]). Because of this heavy usage of social media to share information, users have begun to see it as a channel for obtaining and broadcasting information.[3] A large number of users post about health related information, particularly in

[b] `http://www.statisticbrain.com/twitter-statistic/` [Accessed Sep-29-2015]

online health communities.[16] A recent survey by the Pew Research Center[17] has elucidated the relevance of social media in modern day public health, explaining that 34% of caregivers and 20% of patients read or watch someone elses commentary or experience online. Additionally, 11% of caregivers and 6% of patients share experiences and post questions online. Health related social networks have been attracting many users, perhaps because it allows users of a particular health interest to exchange information. In such platforms (*e.g.*, DailyStrength,[18] MedHelp[19]), users discuss their health-related experiences, including use of prescription drugs, side effects and treatments. Due to the emergence of such platforms, and the abundance of data available through them, research on public health monitoring, including ADR monitoring, has focused on exploiting data from these sources in recent times.[11,20,21]

From the perspective of public health, social media has been utilized for studying smoking cessation patterns on Facebook,[22] identifying user social circles with common medical experiences (like drug abuse),[23] and monitoring malpractice,[24] to name a few. When different patients that suffer from a common disease, or use a specific medication, share information about their symptoms, treatments or drug outcomes, this information can provide valuable clinical insights for both patients and health-related industries that go beyond traditional communication methods.[25] Although specific information about a single user may not be available or usable for privacy reasons, various resources are currently available to perform some demographic analysis with social media data. Furthermore, over the last decade, a number of social media based surveillance systems have been developed, reviewed, and implemented locally, nationally, and globally.[26] The main value of social media, is not derived from individual posts, but from a large number of posts on a specific topic. Recent advances in the data processing capabilities of machines, and machine learning and NLP research present the possibility of utilizing this massive data source for a variety of purposes, including public health. The fact that it is a direct source of users personal experiences makes it a lucrative resource. According to Harpaz *et al.*,[27] social media offers new opportunities for public health monitoring due to the availability of large amounts of data that is internet-based, patient-generated, unsolicited, and up-to-date.

ADR monitoring research have seen significant strides towards the use of automatic NLP techniques for mining drugs and associated reactions from social media. User posts in social media contain information about treatment outcomes and provide early access to reported ADRs that could be beneficial for health and pharmaceutical industries. The type and volume of ADR information that social media makes available to the health industry may not be easily obtainable by other means. This includes the ADRs experienced by those with special conditions, such as patients with rare diseases, pregnant/nursing women, elderly people or patients with co-morbidities who are usually excluded from clinical trials.[28] It is now well established that social media data is rich in knowledge, which is drowned in large volumes of noise.

1.3. *Challenges of social media-based pharmacovigilance*

Various pros and cons of using social media for automatic ADR monitoring,[29,30] and more generally, for public health monitoring, have been mentioned in recent literature. We briefly outline the opportunities that social media presents, and the obstacles associated with its use for health-related research.

The drawbacks found when utilizing the user generated content of social media may include issues with the credibility, recency, uniqueness, frequency, and salience of the data.[3,31] Abbasi and Adjeroh[31] demonstrate the potential downside of each of these five points and the importance of selecting the right media channel for social media analytics. For example, the authors discuss the potential low salience of Twitter because of the short text limits. In addition to these general problems related to the data generated within social media, there are difficulties and challenges posed by the processing and extraction of relevant information using NLP techniques. A frequently encountered challenge is due to the fact that the data is generated by consumers, and they tend to use misspellings, non-medical, descriptive terms to discuss health issues. This reduces a systems ability to automatically extract mentions of relevant concepts and map them to suitable medical lexicons for further analysis.[11,12,15]

Traditional NLP methods that are used on longer texts have proven to be inadequate when applied to short texts, such as those found in Twitter.[32] Thus, recent research tasks have focused on developing NLP tools specifically for data from social media.[33] Some recent articles have reported the imbalance that exists in data coming from social media.[8,10,34] Only a small proportion of drug-associated data collected from social media tend to contain information associated with ADRs. This results in problems associated with annotations, since large volumes of data need to be annotated for the inclusion of sufficient numbers of posts containing ADRs. This data imbalance issue is a major problem for supervised machine learning approaches, particularly because it is the smaller class that is of primary interest for the research. While access to users personal experiences with prescription drugs is one of the key advantages of social media, automatic determination of true personal experiences is challenging. In addition to these, there are also technical, policy, and privacy challenges associated with the use of social media for pharmacovigilance, as pointed out by Edwards and Lindquist.[29]

2. Workshop Task Descriptions

The primary objective of the workshop is to promote the application of different techniques on a common social media based data set, so that useful approaches can be identified and utilized in the future. We divided the overall task of utilizing social media posts for identification of ADR signals into three subtasks:

(1) Automatic classification of ADR assertive user posts (tweets). The goal of this task is to efficiently separate the large amount of noise from posts presenting real ADR associated experiences.
(2) Automatic extraction of ADR mentions. The goal of this task is to apply information extraction techniques to extract text segments so that specific ADRs associated with a drug can be identified.
(3) Normalization of ADR mentions. The goal of this task is to normalize different lexical representations of the same ADR concepts into standard IDs.

To facilitate the shared task, we made available our large annotated Twitter data set. The overall shared task was designed to capitalize on the interest in social media mining and appeal to a diverse set of researchers working on distinct topics such as natural language processing, biomedical informatics, and machine learning. The different subtasks present a number of interesting challenges including the noisy nature of the data, the informal language of the user posts, misspellings, and data imbalance. The rest of this section details the nature of our data and annotations, and each task in detail.

2.1. *Data*

The data set made available for the shared task has been sourced from the social networking site Twitter. The corpus was created through two phases of annotations performed for a large study on ADR detection from social media that is currently in progress. Our finalized annotations are periodically made publicly available at: `http://diego.asu.edu/downloads`.

The tweets associated with the data were collected using the generic and brand names of the drugs, and also their possible phonetic misspellings,[35] since it is common for user posts on Twitter to contain spelling errors. Following the collection of the data, a randomly selected sample of the data was chosen for annotation. The data was annotated by two domain experts under the guidance of a pharmacology expert. Each tweet is annotated for the presence of ADRs (binary), spans of ADRs, indications, and beneficial effects. For each ADR, indication, and beneficial effect, the annotators also identified the most appropriate UMLS concept ID. Following the annotation of the full set, the disagreements were resolved by the pharmacology expert.

2.2. *Task 1: Adverse drug reaction classification*

The first task focuses on automatic classification of ADR assertive user posts. This task utilizes the binary annotations in the data. Participants were provided with a training/development set, containing a set of tweets with associated binary annotations indicating the presence or absence of ADRs. Evaluation was performed on a blind set not released prior to the evaluation deadline. Systems were evaluated on their ability to automatically classify ADR containing posts.

2.2.1. *Training and evaluation sets*

A total of 10,822 annotated tweets were made available [c]. The final data set made available for training is highly imbalanced, as one would expect, with 1,239 (11.4%) tweets containing ADR mentions and 9,583 (88.6%) containing no ADR mentions. Further details about the data set, at an intermediate stage of preparation, and annotations (in addition to the binary annotations) can be found in our past publications.[8,10]

The evaluation set consisted of 4,895 tweets with only 367 (7.4%) ADR instances.

[c] Because of Twitter's privacy policy, the actual tweets cannot be shared. Instead, we have made available a download script and Twitter userIDs and tweetIDs, which interested researchers can use to download the tweets, and associated meta-data.

2.2.2. *Inter annotator agreement*

A randomly chosen subset of the data (1082 tweets) was annotated by the pharmacology expert for the measurement of Inter Annotator Agreement (IAA). We used Cohens Kappa (κ)[36] to compute inter annotator agreement which is given by the following equation. We computed κ for all three pairs of agreements, and obtained an average of 0.71, which can be considered as significant agreement.[37] For the two annotators, $\kappa = 0.69$.

2.3. *Task 2: Adverse drug reaction extraction*

This sub-task is a Named Entity Recognition (NER) task, and the aim is to automatically extract the ADR mentions reported in user posts. This includes identifying the text span of the reported ADRs. Participants were encouraged to use advanced machine learning systems on the annotated training set to extract the mentions and correctly distinguish ADRs from similar non-ADR mentions.

2.3.1. *Training and evaluation sets*

The training data for this sub-task consisted of 2,131 tweets which are fully annotated for mentions of ADR and indications. This set contains a subset of the tweets from task 1 that were tagged as ADR assertive, plus a random set of non-ADR tweets. The non-ADR subset was annotated for mentions of indications, in order to allow participants to develop techniques to deal with this confusion class. To summarize, each instance may contain annotations of medical signs and symptoms with the following semantic types:

- adverse drug reaction– a drug reaction that the user considered negative;
- beneficial effect– an unexpected positive reaction to the drug;
- indication– the condition for which the patient is taking the drug; and
- other– any other mention of signs or symptoms.

Every annotation includes the span of the mention (start/end position offsets), the semantic type, the related drug name, and the corresponding UMLS (Unified Medical Language System) concept ID— assigned by manually selecting concepts in our in-house ADR lexicon.[14] The evaluation set consisted of 476 instances.

2.3.2. *Inter annotator agreement*

We measured inter annotator agreement on the whole training set. The calculated κ value for approximate matching of the concepts is 0.81 for Twitter, which can be considered high agreement.[37]

2.4. *Task 3: Normalization of adverse drug reactions*

This is a concept normalization task. Given an ADR mention in natural language (colloquial or other), participant systems were required to identify the UMLS concept ID for the mention. Unlike the other two tasks, there has not been prior work on normalization of concepts expressed in social media text. We expect immediate future research tasks to focus on this topic.

2.4.1. *Training and evaluation sets*

Training data consists of a set of ADR mentions and their corresponding, human-assigned UMLS concept IDs, as shown below:

`schizophrenia`	`c0036341`
`tension in my nerves`	`c0027769`
`shaking`	`c0040822`

2.5. *Evaluation metrics*

2.5.1. *Task 1 Evaluation*

For this task, the evaluation metric was the ADR F-score. The binary annotation consisted of two classes: ADR and non-ADR. The intent of this task was to devise automatic classification techniques for detecting ADR assertive user posts. As such, the evaluation was based on the harmonic mean of the recall and precision for the ADR class. The ADR F-score has been previously used for evaluation of systems performing this task.[10] The system with the highest ADR F-score on the test set was ranked first.

2.5.2. *Task 2 Evaluation*

F-score was also used as the metric for evaluation in this task. True positives, false positives and false negatives for a system were identified via approximate matching. The F-score was then computed from these values, as described in our past system evaluations.[14]

2.5.3. *Task 3 Evaluation*

For this task, the proposed evaluation metric was accuracy: $\frac{numberofcorrect}{total}$. In this evaluation scheme, a system prediction is considered correct if the predicted concept ID is identical, is a synonym, or has a is-a relationship to the gold standard concept.

3. Methods and Participating Systems

In this section, we summarize the methods used by a selected set of participating teams/systems. We discuss 5 teams' submissions for Task 1 and 2 teams' submissions for task 2. There were no submissions for task 3.

3.1. *Task 1 Systems*

All the submitted systems applied supervised classification approaches. The two best performing systems applied classifier ensembles. The following is a brief discussion of each system.

3.1.1. *Mayo-NLP*

The Mayo-NLP system[38] used an ensemble machine learning classifier to tackle the unbalanced distribution of the classes in the data provided for the task. A feature set containing unigrams, bigrams, and trigrams (a selected list, using mutual information), co-occurrence of drug and side effect, negation, and sentiment score were used to train Random Forest classifiers for identifying ADR assertive tweets. For training, the system obtained best results when the ratio of the training and test sets are balanced, via removal of a random set of negative instances. The system obtained a best F-score of 0.4195.

3.1.2. *TJZZF*

The TJZZF system[39] also uses an ensemble classification strategy. The system uses a weighted average ensemble of four classifiers: (1) a concept-matching classifier based on an ADR lexicon, (2) a maximum entropy (ME) classifier with n-gram features and a TF.IDF weighting scheme, (3) a ME classifier based on n-grams using Naïve Bayes (NB) log-count ratios as feature values, and (4) a ME classifier with word embedding features. This system showed the second best performance with an ADR F-score of 0.4182.

3.1.3. *ReadBioMed*

The READ-BioMed system[40] utilized a few lexical normalization processes and employed existing tools to enrich tweet texts before applying a machine learning-based classifier on the tweets. Unlike the Mayo-NLP system, the focus of this system is to reduce the number of errors caused by the lexical irregularities of tweets. The conceptual enrichment of tweets is based on the sentiment of the tweets, emotion classes, some UMLS Metathesaurus concepts, as well as drug, chemical substance, and disease mentions. The best performance of READ-BioMed on the official test set was achieved using Support Vector Machines (SVMs) trained on a bag-of-words representation for tweets which was enriched with sentiment analysis, emotion classes, and specific UMLS Metathesaurus concepts. The best ADR F-score obtained by this system is 0.358. Importantly, this system shows that enriching text segments via the incorporation of semantic information may be helpful for this task.

3.1.4. *NTTUMUNSW*

The NTTUMUNSW system[41] applied a linear SVM classifier. In addition to n-grams, the system uses a set of lexicon-based features, polarity cues, and topic models derived from the tweets. The best ADR F-score obtained by the system is 0.33. Importantly, the experiments performed using this system show that incorporating features based on topic models improve classification performance.

3.1.5. *SwissChocolate*

The SwissChocolate system[42] adapted a sentiment classification system to the ADR classification task by adding additional features and domain-specific resources. Features include word and character n-grams, POS tags, word clusters and embeddings, and a set of lexicon-based features. The system obtained a relatively low F-score compared to competing systems. However, this system produced very high recall scores.

3.2. *Task 2 Systems*

3.2.1. *DLIR*

The DLIR system[43] uses a very similar technique to the current state-of-the-art in social media based ADR extraction.[14] The system utilizes Conditional Random Fields (CRFs) trained on the annotated data. The system leverages word representations from large amount of unlabeled tweets, both drug related and generic. In addition to using vector representations of words, the system incorporates Part-of-speech tags, n-grams, lexicons, spell-checking, and negations. The best run of the system obtains F-score of 0.611.

3.2.2. *NTTUMUNSW*

The NTTUMUNSW extraction system[44] primarily focuses on token normalization and word representations, and their impacts on extraction. The system utilizes different word representation methods, including token normalization, and two state-of-the-art word embedding methods, namely word2vec and global vectors. The best system run achieved an F-score of 0.540.

4. Results and Discussions

Table 1 presents the results of the different runs of the systems discussed in this paper. The Mayo-NLP-2 system[38] and the TJZZF-1 system[39] achieved the best F-scores, 0.419 and 0.418 respectively. The two runs of the SwissChocolate system[42] performed significantly better than the other systems in terms of recall, but at the cost of precision.

Table 2 presents the results for Task 2. The DLIR system[43]runs significantly outperformed the NTTUMUNSW system runs.[44]

5. Conclusions

The primary aim of this workshop is to facilitate the development of state-of-the-art NLP and machine learning systems that can effectively utilize social media data. We received 11 registrations, of which we have discussed 7 selected system descriptions in this paper. The participating systems explored various interesting properties of social media text, and their impacts on pharmacovigilance oriented tasks.

This is the first time that a shared task is hosted at the Pacific Symposium on Biocomputing 2016. Considering the success of this style of workshop organization, we hope that we will host more of such shared task oriented workshops in the future.

Acknowledgments

Our work on social media mining for adverse drug reaction monitoring is supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) grant number NIH

Table 1. Performances of selected system submissions for the Social Media Mining Shared Task 1. ADR F-scores were used to rank the systems. Best performing system shown in boldface.

System	Precision	Recall	ADR F-score	Accuracy
NTTUMUNSW-1	0.355	0.302	0.327	0.904
NTTUMUNSW-2	0.351	0.244	0.288	0.907
TJZZF-1	0.353	0.512	0.418	0.890
TJZZF-2	0.270	0.578	0.368	0.847
Read-BioMed-1	0.312	0.326	0.319	0.892
Read-BioMed-2	0.358	0.353	0.355	0.901
Read-BioMed-3	0.342	0.371	0.356	0.897
Read-BioMed-4	0.340	0.379	0.358	0.895
Read-BioMed-5	0.358	0.331	0.344	0.903
MayoNLP-1	0.380	0.430	0.403	0.902
MayoNLP-2	0.361	0.501	**0.419**	0.893
MayoNLP-3	0.392	0.408	0.400	0.906
MayoNLP-4	0.431	0.347	0.385	0.914
MayoNLP-5	0.459	0.270	0.338	0.919
SwissChocolate-1	0.202	0.741	0.317	0.754
SwissChocolate-2	0.202	0.743	0.317	0.754

Table 2. Performances of selected system submissions for the Social Media Mining Shared Task 2. ADR F-scores were used to rank the systems. Best performing system shown in boldface.

System	Precision	Recall	ADR F-score
NTTUMUNSW-1	0.782	0.412	0.540
NTTUMUNSW-2	0.718	0.416	0.526
NTTUMUNSW-3	0.778	0.414	0.540
DLIR-1	0.805	0.482	0.603
DLIR-2	0.806	0.485	0.606
DLIR-3	0.760	0.511	**0.611**

NLM 5R01LM011176. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM or NIH.

References

1. C. Kongkaew, P. R. Noyce and D. M. Ashcroft, Hospital Admissions Associated with Adverse Drug Reactions: A Systematic Review of Prospective Observational Studies, in *Annals of Pharmacotherapy*, (7-8)2008.
2. *The Importance of Pharmacovigilance - Safety Monitoring of Medicinal Products* (World Health Organization, 2002).
3. A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya and G. Gonzalez, *Journal of Biomedical Informatics* **54**, 202 (2015).
4. S. Golder, G. Norman and T. K. Loke, *British Journal of Clinical Pharmacotherapy* **80**, 878

(October 2015).

5. J. Parker, Y. Wei, A. Yates, O. Frieder and N. Goharian, A framework for detecting public health trends with twitter, in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13 (ACM, New York, NY, USA, 2013).
6. K. Jiang and Y. Zheng, *Advanced Data Mining and Applications* **8346**, 434 (2013).
7. 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*, 2012.
8. R. Ginn, P. Pimpalkhute, A. Nikfarjam, A. Patki, 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 Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, 2014.
9. A. Patki, A. Sarker, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. O'Connor, K. Smith and G. Gonzalez, Mining Adverse Drug Reaction Signals from Social Media: Going Beyond Extraction, in *Proceedings of BioLinkSig 2014*, 2014.
10. A. Sarker and G. Gonzalez, *Journal of Biomedical Informatics* (2014).
11. 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*, 2010.
12. A. Nikfarjam and G. Gonzalez, Pattern Mining for Extraction of Mentions of Adverse Drug Reactions from User Comments, in *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2011.
13. A. Yates and N. Goharian, ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites, in *Proceedings of the 35th European conference on Advances in Information Retrieval*, 2013.
14. A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn and G. Gonzalez, *Journal of the American Medical Informatics Association (JAMIA)* (2014).
15. K. O'Connor, A. Nikfarjam, R. Ginn, P. Pimpalkhute, A. Sarker, K. Smith and G. Gonzalez, Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions, in *Proceedings for the American Medical Informatics Association (AMIA) Annual Symposium*, 2014.
16. W. Chou, Y. M. Hunt, E. B. Beckjord, R. P. Moser and B. W. Hesse, *Journal of Medical Internet Research* **11**, p. e48 (2009).
17. The Pew Rsearch Center, The social life of health information `http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/`, (2011).
18. Online Support Groups and Forums at DailyStrength `http://www.dailystrength.org`.
19. MedHelp Medical Support Communities `http://www.medhelp.org/forums/list`.
20. C. D. Corley, D. J. Cook, A. R. Mikler and K. P. Singh, *Advances in Computational Biology* (Springer New York, 2010), ch. using Web and Social Media for Influenza Surveillance, pp. 559–564.
21. T. Kass-Hout and H. Alhinnawi, *British Medical Bulletin* **108**, 5 (2013).
22. L. L. Struik and N. B. Baskerville, *J. Med. Internet Res.* **16**, p. e170 (2014).
23. L. C. Hanson, B. Cannon, S. Burton and C. Giraud-Carrier, *J Med Internet Res* **15**, p. e189 (September 2013).
24. A. Nakhasi, R. J. Passarella, S. G. Bell, M. J. Paul, M. Dredze and P. J. Pronovost, Malpractice and Malcontent: Analyzing Medical Complaints in Twitter, in *AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*, 2012.
25. J. Sarasohn-Kahn, The Wisdom of Patients: Health Care Meets Online Social Media `http://www.chcf.org/publications/2008/`

04/the-wisdom-of-patients-health-care-meets-online-social-media, Accessed 29-Sep-2015.

26. D. M. Hartley, *the Milbank Quarterly* **92**, 34 (March 2014).
27. R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu and N. H. Shah, *Drug Safety* **37**, 777 (August 2014).
28. B. H. Stricker and B. M. Psaty, *BMJ* **329** (2004).
29. I. R. Edwards and M. Lindquist, *Drug Safety* **34**, 267 (2011).
30. W. Franzen, *Drug Safety* **34**, p. 793 (2012).
31. A. Abbasi and D. Adjeroh, *Intelligent Systems, IEEE* **29**, 60 (March-April 2014).
32. S. Tuarob, C. S. Tucker, M. Salathe and N. Ram, *Journal of Biomedical Informatics* **49**, 255 (March 2014).
33. O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider and N. A. Smith, Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters, in *Proceedings of the NAACL-HLT*, 2-13.
34. B. W. Chee, R. Berlin and B. Schatz, Predicting Adverse Drug Events from Personal Health Messages, in *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2011.
35. P. Pimpalkhute, A. Patki and G. Gonzalez, Phonetic Spelling Filter for Keyword Selection in Drug Mention Mining from Social Media, in *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2013.
36. J. Carletta, *Computational Linguistics* **22** (1996).
37. A. Viera and J. Garrett, *Family Medicine* **37**, 36 (2005).
38. M. Rastegar-Mojarad, 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.
39. Z. Zhang, J.-Y. Nie and X. Zhang, An ensemble method for binary classificaiton of adverse drug reactions from social media, in *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.
40. B. Ofoghi, S. Siddiqui and K. Verspoor, 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.
41. J. Jonnagaddala, T. R. Jue and H.-J. Dai, 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.
42. D. Egger, F. Uzdilli, M. Cieliebak and L. Derczynski, 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.
43. W. Wang, Mining adverse drug reaction mentions in twitter with word embeddings, in *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.
44. C.-K. Wang, H.-J. Dai, J. Jonnagaddala, T. R. Jue, O. Singh, U. Iqbal and J. Y.-C. Li, NT-TUMUNSW system for adverse drug reactions extraction in Twitter data, in *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.