Data, tools and resources for mining social media drug chatter

Abeed Sarker
Division of Informatics
Department of Biostatistics and Epidemiology
The Perelman School of Medicine
University of Pennsylvania
abeed@upenn.edu

Graciela Gonzalez
Division of Informatics
Department of Biostatistics and Epidemiology
The Perelman School of Medicine
University of Pennsylvania
gragon@upenn.edu

Abstract

Social media has emerged into a crucial resource for obtaining population-based signals for various public health monitoring and surveillance tasks, such as pharmacovigilance. There is an abundance of knowledge hidden within social media data, and the volume is growing. Drug-related chatter on social media can include user-generated information that can provide insights into public health problems such as abuse, adverse reactions, long-term effects, and multi-drug interactions. Our objective in this paper is to present to the biomedical natural language processing, data science, and public health communities data sets (annotated and unannotated), tools and resources that we have collected and created from social media. The data we present was collected from Twitter using the generic and brand names of drugs as keywords, along with their common misspellings. Following the collection of the data, annotation guidelines were created over several iterations, which detail important aspects of social media data annotation and can be used by future researchers for developing similar data sets. The annotation guidelines were followed to prepare data sets for text classification, information extraction and normalization. In this paper, we discuss the preparation of these guidelines, outline the data sets prepared, and present an overview of our state-of-the-art systems for data collection, supervised classification, and information extraction. In addition to the development of supervised systems for classification and extraction, we developed and released unlabeled data and language models. We discuss the potential uses of these language models in data mining and the large volumes of unlabeled data from which they were generated. We believe that the summaries and repositories we present here of our data, annotation guidelines, models, and tools will be beneficial to the research community as a single-point entry for all these resources, and will promote further research in this area.

Keywords:

Social media, data mining, public health, natural language processing, data science.

1 Introduction

In recent years, social media has become a crucial platform for communication, discovery of information, and the sharing of opinions and views [1]. Thus, social media has also emerged as a resource for collecting real-time data directly from public discussions. The social media sphere continues to grow [2], and websites like Twitter attract significant numbers of daily users. Twitter currently has 289,000,000 active users with the number of registered users rising by 135,000 every day [3]. With 58 million tweets per day (9,100 tweets per second), Twitter data is content-rich on everyday

Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016),
pages 99–107, Osaka, Japan, December 12th 2016.

discussions. As a result, Twitter, in addition to other popular social networks, is being actively utilized for a range of tasks including advertising [4], opinion mining [5], political analytics [6], and public health monitoring [7].

From the perspective of public health, systems have been proposed for a variety of tasks including the tracking of the spread of infectious diseases [8, 9], monitoring of prescription and illicit drug abuse [10-12], pharmacovigilance [13], and the monitoring smoking patterns [14]. Despite the obvious use cases for utilizing social media data, national surveillance programs are yet to integrate proposed systems [2]. A prime reason for this are the numerous challenges associated with the use of social media data. While early, keyword-based systems were easily deployable [15], their shortcomings have also been identified [16]. Solving complex natural language processing problems with social media data introduce additional challenges—such as dealing with the use of colloquial language and misspellings [17]. Even data collection from social media faces challenges due to these factors. In addition, the notoriously noisy nature of social media data, and data imbalance hinder system performances [13]. As a result, despite the abundance of health-related knowledge that is encapsulated within the vast social media domain, it is still significantly under-utilized in practical systems.

1.1 Social media and data science

Over the last several years, a flurry of research tasks has successfully employed supervised learning systems that use manually annotated data to solve various natural language processing (NLP) problems. These include, for example, text classification tasks such as detecting mentions of adverse drug reactions [22], and extracting exact mentions using sequence labeling techniques [23]. While these approaches have shown good performance in noisy, social media text, their need for manual annotations make them expensive in nature. Manual annotations are time consuming, and the erratic properties of social media text make annotation tasks even harder. Consequently, even designing annotation tasks and guidelines require significant amounts of expert time, experience in annotations, and exposure to user posted texts. While research from the recent past [13] has elaborated the need for data annotation efforts, the importance of developing standardized annotation guidelines for social media based non-standard data sets have been somewhat overlooked. Therefore, in addition to the need for publicly available targeted data and models, there is also a need for the development of social media text annotation guidelines that to ensure consistency in annotation standards.

The majority of the data available from social media is unlabeled. Recent advances in NLP has seen the effective application of language models learned from large volumes of unlabeled data for various text mining tasks. While the ability to learn language models from large data sets presents new possibilities, social media oriented public health monitoring research has still not actively applied these techniques. One reason behind this is that targeted data from social media for specific public health monitoring tasks is still scarce. Thus, there is a strong motivation for the public release of such data and models. For example, recent approaches for generating distributed word representations [18-20] from large, unstructured data sets have seen growing popularity. However, availability of such language models learned from relevant social media data is limited.

1.2 Aims

We have several aims for this broad coverage paper. These aims are summarized as follows:

1. To outline our annotated data and the resources we have created over the last several years, as part of a National Institute of Health (NIH) research grant [21] on mining social media for discovering adverse drug reactions.
2. To make available our evolving social media text annotation guidelines for pharmacovigilance and toxicovigilance so that these annotation guidelines can be followed for future annotation tasks.
3. To provide insights about our annotated corpora, annotation tasks, unlabeled data and models.
4. To discuss some of the utilities of our data sets and their potential future uses.

The rest of the paper is organized as follows. In the next section we present (i) our data collection technique, which expands on keyword-based approaches to include common, phonetically similar misspellings of drug names, (ii) our preparation of various publicly available annotated data sets, (iii) our detailed annotation guideline preparation, and (iv) our language model generation techniques. In the *Discussion* section we present some statistics and utilities of our published resources and tools, including potential applications of our unlabeled data and language models.

2 Methods

2.1 Data collection

Prior to collecting data, we selected a set of drugs of interest, which were likely to have a large number of associated comments in social media. In particular, we selected drugs that were prescribed for chronic diseases and syndromes for which large numbers of comments were expected and drugs with high prevalence of use (as per the IMS Health's Top 100 drugs by volume for the year 2013 [22]). Starting with this initial list of drugs, we added various drug names based on interest since 2014, such as drugs that may be prone to abuse. The final drug list is monitored by our in-house pharmacology expert, and further details about the drugs can be found in our past publications [22,23].

We collected data from Twitter using the drug names (trade and generic) as keywords. To address the issue of misspelled drug names, which affects recall during data collection, we developed a spelling variant generator [24]. The generator first identifies lexically close misspellings, specifically those that are 1-edit distance away in terms of Levenshtein distance. Phonetically similar misspellings are then identified, and finally, the Google custom search API is used to identify a smaller set of misspellings that are commonly used by users. We have made a downloadable version of our generator publicly available.[1] The generator is semi-automatic. Figure 1 presents a random sample of tweets associated with a number of drugs that were collected using our technique. The tweets appear to present a number of types of information, such as symptoms/indications, perceived adverse drug reactions, medication abuse information, user sentiments towards drugs and/or prices, and potential drug abuse, to name a few. The figure also illustrates how some drug names are often misspelled. Depending on the intent, distinct types of drug-related information can be mined from this data source.

2.2 Data annotation, guidelines and resources

Following the collection of large amounts of drug-related chatter from Twitter, we allocated significant resources to perform annotation of the data and for the preparation of standardized annotation guidelines. The annotation guidelines were prepared in consultation between experienced language annotators, NLP experts, public health professionals, and a pharmacology expert. The guidelines were finalized by the pharmacology expert after multiple iterations. The annotation guidelines also evolved over time, which is a necessity for social media data, as new characteristics of the data were discovered during the early iterations of annotation. Using the annotation guidelines, we were able to achieve high inter annotator agreements for our various annotation tasks. For adverse drug reaction detection from social media, we first performed binary annotations indicating if user posts mentioning at least one drug mentioned an adverse reaction or not (inter annotator agreement $\kappa = 0.74$). Following that, we performed annotations to tag specific mentions of adverse reactions and indications ($\kappa = 0.81$), including mapping the mentions to standardized IDs in the Unified Medical Language System (UMLS) vocabulary. We have made these detailed annotation guidelines publicly available to support future annotation tasks.[2] In addition to the guidelines, we have made resources associated with our classification and extraction tasks publicly available [22,23]. These include source codes, executable applications, lexicons, topics, cue words, word clusters, word embeddings, and annotated data, which we discuss later.

[1] Available at: http://diego.asu.edu/Publications/ADRSpell/ADRSpell.html. Last accessed: 2nd October, 2016.

[2] Available at: http://diego.asu.edu/guidelines/adr_guidelines.pdf. Last accessed: 2nd October, 2016.

can't sleep, **temazepam** myself into a coma, pass out for hours on end. finally wake up, feel like shite for days. Oh I love my life! :-/

my fibromyalga is killing me lately. has anything worked for u? **lyrica** and **neurontin** f'd up my life. **cymbalta** worse

just got retested for jcv. **tecfidera** did not work out well for me, so i'm onto **tysabri**. #ms #multiplesclerosis

adderal made me manic, **saphris** makes my skin crawl and gives me the dreaded twitches, **hydroxyzine** is more like a placebo than anything else

list of psychiatric medications i take for various psychiatric reasons. !. **saphris**. 2. **lamictal**. 3. **hydroxygine**. 4. **trazodone**. 5. **zoloft**.

the only kind i have is sleeping **siroquil** and it knocks me out for too long to make it to class

the sun is up & i haven't slept yet! the **quetiapine** is not knocking me out like it used to. been up for 24 hours & i aint sleepy :-(#bipol

snorted 2 15mg **oxycodone** ($24)

also **adderall** prevents me from having any feelings other than tired rage

i hate how this firbo and **gabapentin** robs me if my life ... i just hate feeling so useless and worthless feeling tired

i am taking a cocktail of **tramadol**, **acoxia**, **myonol** & **pregabalin** twice a day and I still cannot control this pain. huhuhuhuhuh

do not take **victoza** if you are allergic to **victoza** ... i an now worried about people who actually need this warning

i'm trying to go off it. i'm on **lamictal** now and it works but i'm still addicted to **Geodon**

my memory is still so awful, hate the side effects of **pregabalin** -.-

Figure 1. Sample tweets containing drug names including some that are misspelled, but were caught our common misspelling generator. The tweets present a variety of different types of information including the symptoms effectiveness of drugs, adverse reactions, user sentiments, and potential abuse of prescription medications.

In addition to our work on pharmacovigilance, our experts have collaborated to create guidelines and resources for additional tasks such as prescription medication abuse monitoring from social media. Similar to our other tasks, the annotations were carried out in several iterations and the guidelines prepared have been made publicly available.[3] We have also made some of our annotated research data on peripheral topics, such as prescription medication abuse, publicly available.[4]

As discussed in the abovementioned guidelines, annotation of social media data presents a variety of challenges, which must be addressed in a consistent manner. For Twitter, the first challenge faced when performing binary annotations was the lack of context. Due to the character limit of 140 per post, even for human annotators, it is often difficult to determine the context in which a potential adverse reaction is mentioned or if a mentioned adverse reaction represents a personal experience or just a general statement. Other factors, such as posts that are spread over multiple tweets also add to this problem. To address these and other annotation difficulties, regular meetings were held between the annotators and the pharmacology expert, during which common difficult annotation issues were identified, discussed, and resolved. We provide further details of common social media text annotation problems that we faced in the *Discussion* section.

2.3 Unlabeled data and language models

Besides preparing and releasing the largest annotated social media data sets for pharmacovigilance and other tasks, we also released unlabeled data and language models derived from the data. Language models generated from unstructured data sets, such as those via deep learning techniques, have recently received significant research attention because of their ability to capture semantic information [18]. We released two sets of language models for the research community, along with the data (approximately a quarter million tweets) used to create the models.[5] The following is a brief overview of each set.

The first set of models were prepared using the word2vec tool,[6] and they capture distributional and semantic information. Phrases/terms are represented using vectors using these models, with the vector sizes largely determining where each phrase appears in semantic space. We generated models with vector sizes between the sizes 200 and 400. For the different vector sizes, we generated models using context windows within the range [2,9]. Such distributed word representation models are already being applied for research utilizing other sources of noisy health-related data, such as clinical reports [25]. Our second set of models are sequential, and these language models capture the probabilities of n-gram sequences. These models have been applied for a variety of tasks in the past, such as lexical normalization [26]. In a sequential language model, the conditional probability of a term given all the previous terms is given as $P(t_1^M) = \prod_{k=1}^{M} P\left(t_k | t_1^{k-1}\right)$, where t_k is the k^{th} term. To generate the n-gram language models, we used the KenLM n-gram, language modeling tool [27]. We have also made available a set of n-gram language models (n= 2—4) from the same unlabeled data set.

3 Discussions

In this section, we briefly discuss some of the uses of the various resources that we have published. The value of most of our various annotated data sets has already been established, and there has been a sizable amount of recent research that have utilized these data sets for tasks such as classification and extraction. The resources associated with our annotated data sets, such as the lexicons, word clusters, and so on, have been used for research outside the domain of pharmacovigilance. For binary classification of adverse drug reaction classification, we currently have a total of 25,678 annotated posts, which were prepared in 3 batches. 10,822 posts were made publicly available with our system/source code for social media text classification for pharmacovigilance [22].[7] Additional data sets

[3] Available at: http://diego.asu.edu/guidelines/DrugAbuseAnnotationGuideline1.1.pdf. Accessed 2nd October, 2016.

[4] Available at: http://diego.asu.edu/Publications/DrugAbuse_DrugSafety.html. Accessed 2nd October, 2016.

[5] Available at: http://diego.asu.edu/Publications/Drugchatter.html. Accessed 2nd October, 2016.

[6] https://www.tensorflow.org/versions/r0.11/tutorials/word2vec/index.html. Accessed: 20th October, 2016.

[7] Resources, tools and data are available at: http://diego.asu.edu/Publications/ADRClassify.html. Accessed: 26th October, 2016.

were made available to the participants of a shared task that we organized [33], and these data sets will also be made available via the link mentioned above. For adverse drug reaction mention extraction, we have made available 1784 annotated posts publicly available along with our state-of-the-art extraction system [23].[8] In total, we have 2607 annotations for this task, with the rest of the data only available to our shared task participants and will be made publicly available in the near future. We have also made available a collection of resources for social media mining for pharmacovigilance along with our review of the domain [13].[9]

Annotating biomedical data or social media data are challenging tasks and require expertise with the domains. The challenges are exacerbated when it comes to biomedical data from social media. As mentioned in the *Methods* section, we faced several frequently occurring annotation difficulties, which had to be resolved via multiple meetings and paired annotation sessions. The lack of context available with the short Twitter posts often made it difficult to determine if a post mentioned a personal experience of adverse reaction or just mentioned an adverse reaction for other reasons (*e.g.*, in many posts we found users simply repeating adverse reactions mentioned in television commercials). In many cases, our annotators found it difficult to determine if a mentioned condition was an adverse reaction or a symptom for which the drug in question was taken. Annotating the spans of concept mentions is even more challenging. Non-standard expressions (*e.g.*, '*head feels like a zombie'*) and disjoint mentions of adverse reactions (*e.g., 'gives me pain in my freakin stomach*) are two of the leading causes of these difficulties. In addition to annotating the spans, our annotators were also required to map them to IDs in the UMLS. Non-standard adverse reaction mentions and context ambiguities led to numerous cases where more than one concept ID seemed valid. To resolve difficulties in selecting concept IDs, our annotators used paired annotation to identify IDs that were the most concrete fits, and developed specific, step-wise rules which are detailed in the previously mentioned annotation guideline.

Because of the costs and difficulties faced when annotating data within this complex domain, the preparation of comprehensive guidelines, such as ours, is of paramount importance. Detailed annotation guidelines with specific examples of problem cases can significantly reduce time required to plan for and design annotation tasks for social media based NLP studies. Even within the same annotation task there are inconsistencies in distinct research groups. We discovered such inconsistencies, for example, in the several data sets for binary classification of adverse drug reaction mentions. Therefore, we believe that our publicly available annotation guidelines will be helpful for the better understanding of potential issues associated with annotation of social media health data and to plan future annotation tasks.

We have discussed the recent release of a small batch of unlabeled data and sets of language models that were prepared using this data [32]. Analysis of that batch of unlabeled data revealed that discussions associated with drugs are generally skewed in Twitter, with some drugs discussed much more frequently than others. In the abovementioned sample of unlabeled data, while the distribution of tweets over the months were similar, we found some drugs to have a very large number of tweets associated with them. Figure 2 illustrates this information, showing that among the discussions regarding the top 10 most discussed drugs, 56% of the discussion was about Adderall® and 12% was about Xanax®. We suspect that the skewness in the distribution of drugs in social media chatter is because of the demographics among which social media is popular. Adderall®, for example, is a popular medication for abuse among young students, and, therefore, there is a large amount of chatter available for this drug, particularly during typical college examination times (*e.g.*, November/December) [10,28].

[8] Resources, tools and data are available at: http://diego.asu.edu/Publications/ADRMine.html. Accessed: 26th October, 2016.

[9] Available at: http://diego.asu.edu/Publications/ADRSMReview/ADRSMReview.html. Accessed: 26th October, 2016.

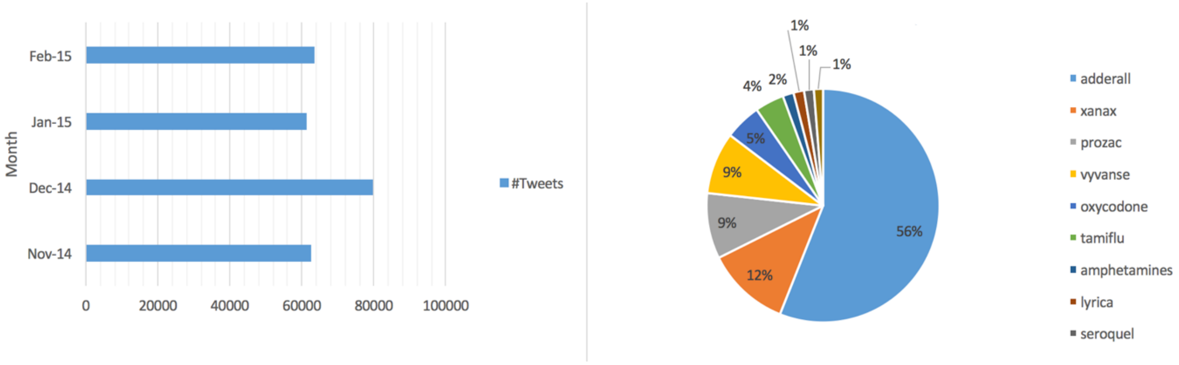
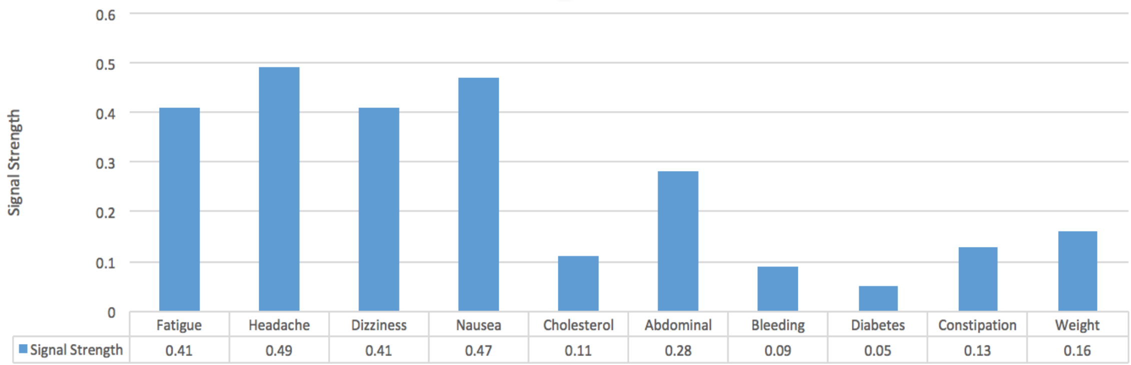


Figure 2. Distribution of drug related tweets over time and over different drug related keywords.

Past research has explored co-occurrence based techniques for identifying drug-adverse reaction associations [29]. One of the properties of the distributional semantics model is the ability to capture semantic associations between terms based on co-occurrence, and we have performed preliminary experimentation to assess the use of our models for drug-adverse reaction association identification. For the drug Trazodone, using one of our distributed representation models with a vector size of 400 and context window size 9, we compared the cosine similarity values between a drug keyword and a set of adverse reaction terms. Our similarity computations produced relatively high scores for known adverse reactions (the first four reactions from the left in Figure 3) and low scores for reactions for which no associations are known. While the threshold for this drug appears to be between 0.3 and 0.4, we could not establish specific values during our preliminary experimentation. Experimentation with other drugs (*e.g.*, such as those presented in [32]), also suggest that thresholds may vary between drugs or classes of drugs. Furthermore, there are unsolved NLP based problems, such as the vector representation of multi-word adverse reaction expressions. We also performed preliminary experiments with our sequential language models, such as assessing their usage in text classification. Because our data set essentially consists of health-related tweets, we used the sequential models to score a sample of posts from a separate data set containing annotations for health related tweets [31]. We observed that in general, health-related posts obtained higher scores compared to non-health related posts, as was expected. However, as with the distributional language models, we could not identify thresholds in the preliminary experimentation. We plan to address some of these limitations of our work in future research. We believe that incorporation of information from these models will improve the existing tasks of classification and extraction, and will be crucial for previously unexplored tasks such as concept normalization.



	Fatigue	Headache	Dizziness	Nausea	Cholesterol	Abdominal	Bleeding	Diabetes	Constipation	Weight
Signal Strength	0.41	0.49	0.41	0.47	0.11	0.28	0.09	0.05	0.13	0.16

Figure 3. Association between trazodone and 10 adverse reactions computed using the distributed language models and cosine similarity.

The experimental results obtained from the use of our language models are very promising. With very simplistic settings, there appears to be a clear use case for these models for the tasks discussed. Our planned future work involves in-depth exploration of the various parameters of these models (*e.g.*, effect of context window sizes). We also encourage the research community to investigate the properties of the distinct models we are making available, and their applications. As discussed earlier, studies have already focused on extracting drug abuse information from social media, assessing the safety of drugs, exploring the prevalence of use of drugs, and discovering user sentiments towards specific drugs, to name a few. The linguistic regularities and the semantic knowledge captured by these models are likely to be useful for a number of important research tasks.

With the ever growing size of social media data, and the development of more efficient data processing techniques, the broader health domain will invariably benefit from utilizing social media data. However, it has also been realized that the *right* data is more important than *big* data, and the development of effective systems benefit from access to the former. Therefore, we believe that our released data, tools and resources, which have been summarized in this paper, will be very useful to the research community.

References

1. Horvitz E, Mulligan D. Data, privacy, and the greater good. Science 2015; 349 (6245):253—255. PMID: 26185242.
2. Velasco E, Agheneza T, Denecke G, et al. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q 2014; 92 (1): 7—33. PMID: 24597553.
3. Web reference. WebCite archive: http://www.webcitation.org/6ceI98x1Z. Original URL: http://www.statisticbrain.com/twitter-statistics/.
4. Khang H, Ki E-J, Ye L. Social Media Research in Advertising, Communication, Marketing, and Public Relations, 1997—2010. Journal Mass Commun 2012; 89 (2): 279—298. DOI 10.1177/1077699012439853.
5. Hu M, Liu B. Mining and summarizing customer review. Proceedings of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data Mining; 2004 Aug 22—25; Seattle, Washington. 168—177. ACM; 2004.
6. Eom Y-H, Puliga M, Smailović et al. Twitter-Based Analysis of the Dynamics of Collective Attention to Political Parties. PLoS One 2015; 10 (7). DOI 10.1371/journal.pone.0131184.
7. Denecke K, Krieck M, Otrusina L et al. How to Exploit Twitter for Public Health Monitoring. Methods Inf Med 2013; 52 (4): 326—339. PMID: 23877537.
8. Paul MJ, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. Proceedings of the International AAAI Conference on Weblogs and Social Media: 2011 July 17—21; Barcelona, Spain. AAAI Press; 2011.
9. Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. PLoS One 2013; 8 (12): e84672. PMID: 24349542.
10. Hanson CL, Burton SH, Giraud-Carrier C et al. Tweaking and tweeting: Twitter for nonmedical use of psychostimulant drug Adderall among college students. J Med Internet Res 2013; 15 (4): e62. PMID: 23594933.
11. Hanson CL, Cannon B, Burton S et al. An Exploration of Social Circles and Prescription Drug Abuse through Twitter. J Med Internet Res 2013; 15 (9): e189. PMID: 24014109.
12. Cavazos-Rehg, Krauss M, Fisher SL et al. Twitter chatter about marijuana. J Adolesc Heal 2015; 56 (2): 139—145. PMID: 25620299.
13. Sarker A, Ginn R, Nikfarjam A et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform 2015; 54: 202—212. PMID: 25720841.
14. Struik LL, Baskerville NB. The role of Facebook in Crush the Crave, a mobile- and social media-based smoking cessation intervention: qualitative framework analysis for posts. J Med Internet Res 2014; 16 (7): e170. PMID: 25016998.

15. Cook S, Conrad C, Fowlkes A et al. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS One 2011; 6 (8): e23610. PMID: 21886802.

16. Lazer D, Kennedy R, King G et al. The Parable of Google Flu: Traps in Big Data Analysis. Science 2014; 343 (6176): 1203—1205. PMID: 24626916.

17. Leaman R, Wojtulewicz L, Sullivan R et al. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reaction from User Posts to Health-Related Social Networks. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing: 2010 July 15; Uppsala, Sweden. 117—125. Association for Computational Linguistics; 2010.

18. Mikolov T, Chen K, Corrado G et al. Efficient Estimation of Word Representations in Vector Spaces. Proceedings of the Workshop at the International Conference on Learning Representations: 2014 May 2—4; Scottsdale, Arizona. Archived at: **arXiv:1312.5650v3.**

19. Mikolov T, Sutskever I, Chen G et al. Distributed Representations of Words and Phrases and their Compositionality. Proceedings of the Twenty-seventh Annual Conference on Neural Information Processing Systems: 2013 December 5—10; Lake Tahoe, Nevada. Curran Associates, Inc; 2013.

20. Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations. Proceedings of NAACL-HLT: 2013 Jun 9—14; Atlanta, Georgia. 746—751. Association for Computational Linguistics; 2013.

21. NIH Grant Number 5R01LM011176, Mining Social Network Postings for Mentions of Potential Adverse Drug Reactions. 2012. RePORT URL: https://projectreporter.nih.gov/project_info_description.cfm?projectnumber=5R01LM011176-02.

22. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015; 53: 196—207. PMID: 25451103.

23. Nikfarjam A, Sarker A, O’Connor K et al. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Informatics Assoc 2015; 22 (3): 671—681. PMID: PMID: 25755127.

24. Pimpalkhute P, Patki A, Nikfarjam A. Phonetic Spelling Filter for Keyword Selection in Drug Mention Mining from Social Media. AMIA Jt Summits Transl Sci Proc 2014; 2014: 90—95. PMID: 25717407.

25. Henriksson A, Kvist M, Dalianis H et al. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform 2015; 57: 333—349. PMID: 26291578.

26. Han B, Baldwin T. Lexical normalization of short text messages: makn sens a #twitter. Proceedings of ACL-HLT: 2011 Jun 19—24; Portland, Oregon. 368—378. Association for Computational Linguistics; 2011.

27. Heafield K, Pouzyrevsky I, Clark JH et al. Scalable Modified Kneser-Ney Language Model Estimation. Proceedings of ACL: 2013 Aug 4—9; Sofia, Bulgaria. 690—696. Association for Computational Linguistics; 2013.

28. Sarker A, O’Connor K, Ginn R, Scotch M, Smith K, Malone D, Gonzalez G. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. Drug Saf. 2016 Mar;39(3):231-40. doi: 10.1007/s40264-015-0379-4. PMID: 26748505.

29. Nikfarjam A, Gonzalez G. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. AMIA Annu Symp Proc 2011; 2011: 1019—1026. PMID: 22195162.

30. Toutanova K, Moore RC. Pronunciation modeling for improved spelling correction. Proceedings of ACL: 2002 Jul 7—12; Philadelphia, Pennsylvania. 144—151. DOI: 10.3115/1073083.1073109. Association for Computational Linguistics; 2002.

31. Paul MJ, Dredze M. A Model for Mining Public Health Topics from Twitter. Technical Report. Johns Hopkins University 2011. Archived at: http://www.cs.jhu.edu/~mpaul/files/2011.tech.twitter_health.pdf.

32. Sarker A, Gonzalez G. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. [To Appear]. Data Brief. 2016.

33. Sarker A, Nikfarjam A, Gonzalez G. Social Media Mining Shared Task Workshop. Pac Symp Biocomput. 2016;21:581—592. PMID: 26776221.