Journal of Biomedical Informatics 53 (2015) 196–207

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Portable automatic text classification for adverse drug reaction detection via multi-corpus training

Abeed Sarker *, Graciela Gonzalez

Department of Biomedical Informatics, Arizona State University, 13212 East Shea Blvd., Scottsdale, AZ 85259, USA

ARTICLE INFO

Article history:
Received 7 July 2014
Accepted 2 November 2014
Available online 8 November 2014

Keywords:
Pharmacovigilance
Adverse drug reaction
Social media monitoring
Text classification
Natural language processing

Objective: Automatic detection of adverse drug reaction (ADR) mentions from text has recently received significant interest in pharmacovigilance research. Current research focuses on various sources of text-based information, including social media—where enormous amounts of user posted data is available, which have the potential for use in pharmacovigilance if collected and filtered accurately. The aims of this study are: (i) to explore natural language processing (NLP) approaches for generating useful features from text, and utilizing them in optimized machine learning algorithms for automatic classification of ADR assertive text segments; (ii) to present two data sets that we prepared for the task of ADR detection from user posted internet data; and (iii) to investigate if combining training data from distinct corpora can improve automatic classification accuracies.

Methods: One of our three data sets contains annotated sentences from clinical reports, and the two other data sets, built in-house, consist of annotated posts from social media. Our text classification approach relies on generating a large set of features, representing semantic properties (*e.g.*, sentiment, polarity, and topic), from short text nuggets. Importantly, using our expanded feature sets, we combine training data from different corpora in attempts to boost classification accuracies.

Results: Our feature-rich classification approach performs significantly better than previously published approaches with ADR class *F*-scores of 0.812 (previously reported best: 0.770), 0.538 and 0.678 for the three data sets. Combining training data from multiple compatible corpora further improves the ADR *F*-scores for the in-house data sets to 0.597 (improvement of 5.9 units) and 0.704 (improvement of 2.6 units) respectively.

Conclusions: Our research results indicate that using advanced NLP techniques for generating information rich features from text can significantly improve classification accuracies over existing benchmarks. Our experiments illustrate the benefits of incorporating various semantic features such as topics, concepts, sentiments, and polarities. Finally, we show that integration of information from compatible corpora can significantly improve classification performance. This form of multi-corpus training may be particularly useful in cases where data sets are heavily imbalanced (*e.g.*, social media data), and may reduce the time and costs associated with the annotation of data in the future.

1. Background

Early detection of adverse drug reactions (ADRs) associated with drugs in their post-approval periods is a crucial challenge for *pharmacovigilance* techniques. Pharmacovigilance is defined as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug problem" [1]. Due to the various limitations of pre-approval clinical trials, it is not possible to assess the consequences of the use of a particular drug before it is released [2]. Research has shown that adverse reactions caused by drugs following their release into the market is a major public health problem: with deaths and hospitalizations numbering in millions (up to 5% hospital admissions, 28% emergency visits, and 5% hospital deaths), and associated costs of about seventy-five billion dollars annually [3–5]. Thus, post-marketing surveillance of drugs is of paramount importance for drug manufacturers, national bodies such as the U.S. Food and Drug Administration (FDA), and international organizations such as the World Health Organization (WHO) [6]. Various resources have been utilized for the monitoring of ADRs, such as voluntary reporting systems and electronic health records. The rapid growth of electronically available health related

* Corresponding author.
E-mail addresses: abeed.sarker@asu.edu (A. Sarker), graciela.gonzalez@asu.edu (G. Gonzalez).

http://dx.doi.org/10.1016/j.jbi.2014.11.002
1532-0464/© 2014 The Authors. Published by Elsevier Inc.

information, and the ability to process large volumes of them automatically, using natural language processing (NLP) and machine learning algorithms, have opened new opportunities for pharmacovigilance. In particular, annotated corpora have become available for the task of ADR identification in recent times, making it possible to implement data-centric NLP algorithms and supervised machine learning techniques that can aid the detection of ADRs automatically [2].

One domain where data has grown by massive proportions in recent years, and continues to grow, is social media [7]. In addition to generic social networks (*e.g.*, Twitter), those focusing specifically on issues related to health have also been attracting many users. In such platforms, users discuss their health-related experiences, including the use of prescription drugs, side effects and treatments. Users tend to share their views with others facing similar problems/results, which makes such social networks unique and robust sources of information about health, drugs and treatments. One such social network (also referred to as online health community), dedicated to health related discussions, is DailyStrength.[1] Due to the emergence of such social media, and due to the abundance of data available through them, ADR detection research in recent times has focused on exploiting data from these sources [8]. Social media based data sources, however, also present various NLP challenges. For example, it has been shown in past research that automated systems frequently underperform when exposed to social media text because of the presence of novel/creative phrases and misspellings, and frequent use of idiomatic, ambiguous and sarcastic expressions [8]. In addition, when attempting to utilize social media data for ADR monitoring, problems of data imbalance and noise are introduced. Fig. 1 illustrates several examples of social media posts exhibiting the abovementioned problems. The posts express the users' views about specific medications. It can be observed from the figure that there are frequent misspellings (*e.g.*, 'seroquil', 'numbb', 'effexer', 'bfore'), use of ambiguous/non-standard terms for expressing adverse reactions (*e.g.*, 'look like a zombie', 'ton of weight'). These properties of the texts hamper the identification and generalization of the lexical properties of different posts, thus, adversely affecting the performance of automatic rule-based and learning-based approaches. The problem is further exacerbated by the fact that the posts are generally very short, and so only limited features can be extracted via shallow processing.

Using advanced NLP techniques and resources, deep semantic and linguistic features can be extracted from these texts. These features can be used to indicate mutually exclusive properties of the posts along various dimensions (*e.g.*, sentiment, polarity, topic, *etc.*). The various properties may then be combined to generate a rich set of features for each post, thus, aiding the process of automatic classification, and consequently, automatic detection of ADRs. In this paper, we address the problem of automatic detection of ADR assertive text segments from distinct sources, particularly focusing on user posted data. This automatic text classification mechanism forms a crucial component of an automatic, social media-based ADR detection/extraction pipeline. This component is essential to the ADR monitoring system because most of the data from social media is irrelevant for the task of ADR detection, and must be filtered out before the data is processed by modules responsible for other tasks. Fig. 2 illustrates our pipeline to detect ADRs from social media at a very high level. In this paper, we discuss steps 1 and 2, with particular focus on the second step.

Because of the popularity of social networks, and their high growth rates, they promise to be very lucrative sources of information which can be utilized for pharmacovigilance tasks. Currently, there is limited research that attempts to apply advanced NLP techniques to extract features from user-posted text for ADR detection (*e.g.*, NLP-based sentiment analysis techniques [9]). There is also no work on combining data from multiple social networks to improve ADR detection/classification performance. Furthermore, there are no large publicly available corpora[2] for research or comparison of systems. Thus, there is a strong motivation behind the research we present in this paper. We discuss two social networks from which we prepare our annotated data (one of the data sources have been made publicly available). Since social media posts are generally short, we attempt to generate features representing various properties of the texts to increase the number of features that can be learnt. We focus on the application of NLP approaches to extract indicative features from text and utilize supervised machine learning techniques to automatically classify text segments indicating ADRs. Finally, we combine data from multiple social media sources in an attempt to improve classification accuracies.

1.1. Intent

Our primary intent is to explore the extent to which ADR assertive text segments can be automatically classified from text-based data sources, particularly social media sources. We are interested in exploring if text from various sources can be combined to improve classification accuracies, specifically in the presence of imbalanced data. The following list summarizes our intents:

(i) To explore NLP techniques which can be used to extract informative and portable features from text coming from distinct sources, including social media. In particular, we experiment with some novel features and some previously proposed features that have not been applied for the task of ADR detection.
(ii) Investigate the performance of supervised classification approaches on data from social media compared to data from other more structured sources.
(iii) Investigate approaches by which the data imbalance problem associated with automatic ADR detection from social networks can be overcome, and optimize machine learning algorithms to improve performance over existing approaches.
(iv) Investigate the possibility of combining annotated data from multiple sources to boost automatic ADR classification accuracies.

1.2. Contributions

The contributions we make in this paper are as follows:

(i) We show how NLP techniques can be applied to extract useful features from text that can improve classification performance over existing approaches. We propose novel features and import several useful features from similar text classification research. We provide descriptions of the extracted features along with the intuitions behind their use, and an analysis of their impacts.
(ii) We compare the differences in performances of machine learning algorithms over data sets from distinct origins.
(iii) We address the issue of data imbalance for the ADR detection task. We show how multi-corpus training can help alleviate the problem of data imbalance and further improve the performance of ADR detection.
(iv) We present a data set developed in-house and made publicly available for research purposes.

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

[2] We identified two very small corpora that are publicly available. They are discussed in the next section.

Twitter

- HA! Not if you're on #Seroquil. EXTREMELY vivid dreams that stay in conscious memory. Very #Freaky! Any idea why?
- Seriously if you are getting off cymbalta tape alot if movies I was up till 4 and feel and look like a zombie just to tired to growl !!!
- I'd rather be on the Effexor that made me happy though I had severe back pain from it.. Than sad and crying for no reason on lexapro....
- -nods- My zombie-ness when I first wake up in the morning is mostly from my nightly meds too (Seroquel).
- but first! Try these lovely pharmies! #zoloft feel numbb #paxil hate life more and everyone else

DailyStrength

- Don't like the weight gain that comes with it.
- My husband hates this med. Gives him shakes and tremors. He also has gained a ton of weight and now has acid reflux etc.
- Was horrible experience. Spiraled her into suicidal state, ended up in hospital
- knocks me out cold but i hate the feeling right bfore falling asleep
- my step-dad was on Effexer,taking supplements for energy and drinking like a fish when he shot my daughter and me

Fig. 1. Examples of user posts on Twitter and DailyStrength regarding their experiences with drugs. User names have been removed from the posts for privacy protection.

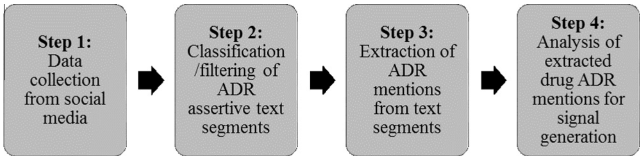
Step 1: Data collection from social media
Step 2: Classification /filtering of ADR assertive text segments
Step 3: Extraction of ADR mentions from text segments
Step 4: Analysis of extracted drug ADR mentions for signal generation

Fig. 2. ADR detection pipeline from a very high level.

For the experiments described in this paper, we utilize data from three different sources: one from a generic social network (Twitter), one from an online health community (DailyStrength), and the third from clinical reports (the ADE corpus [10]). We provide further details of the data in Section 3. We provide an overview of related work in this area in Section 2 and discuss our approaches in detail in Section 4. In Section 5, we present comparative evaluations of the performance of our approach with earlier approaches, and over distinct data sets. We also discuss the contributions of various features, and the primary sources of classification errors. We conclude the paper in Section 6, and discuss potential future work.

2. Related work

2.1. Automatic ADR detection

In the past, drug safety surveillance has primarily relied on spontaneous reporting systems. These are passive systems populated by reports of suspected ADRs collected from healthcare professionals, pharmaceutical companies, and maintained largely by regulatory and health agencies [2]. One such system is the Adverse Event Reporting System (AERS) of the U.S. Food and Drug Authority (FDA). However, recent research suggests that AERS grossly under estimates the prevalence of serious ADRs [11]. Such spontaneous reporting systems also suffer from a range of limitations including under-reporting (only approximately 10% of serious ADRs are reported [12] in AERS), over-reporting of known ADRs, incomplete data, duplicated reporting and unspecified causal links [13,14]. Thus, in recent times, research focus has broadened to the utilization of other sources of data for ADR detection.

NLP approaches have been proposed to detect ADRs from text, and they have mostly focused on utilizing data from Electronic Health Records (EHR) [15–17], and also other sources such as clinical reports [6,10]. Unlike spontaneous reports, electronic health records contain more complete records of patients' medical history, treatments, conditions, and potential risk factors, and are also not limited to patients who experience ADRs [18]. Thus, they have several advantages over spontaneous reporting systems. Electronic health records, however, have their own challenges, such as that of pervasiveness and confounding, and the definition and ascertainment of exposures and outcomes [2].

Due to the limitations of a single source of information, some recent research has focused on combining information from multiple sources such as the combination/enrichment of information gathered from spontaneous reporting systems with literature findings [19–21], electronic health records [18], and chemical similarity [22], and even user posted data on social media [21]. It has been shown that combining information from multiple sources can significantly boost ADR detection accuracies. Providing an elaborate review of all ADR detection approaches is outside the scope of this paper. We recommend the interested reader to refer to Harpaz et al. [2] for a detailed review of ADR mining techniques from various sources.

2.2. ADR detection from text and social media

Social media data presents interesting opportunities and unique challenges to NLP techniques for ADR detection. Social networks contain a large volume of user posted data and are increasingly being used as tools for real-time knowledge discovery [23], but the presence of non-standard and colloquial terms and various other factors make the task of data mining harder [24]. In addition, the large volume of irrelevant information that must be filtered out, as mentioned earlier, makes the utilization of the data further difficult. However, despite these difficulties, social media data is a rich and lucrative source of information. Social networks on health related topics have seen rapid growth in recent years and it has been observed that users often share health experiences with their peers rather than during clinical research studies or with their physicians [25]. Despite the vast amount of information available on social networks, research on mining that data for ADR detection is still very much in its infancy. The earliest work to utilize user comments was that by Leaman et al. [8], who showed that user comments from health-related social networks can reflect known ADRs and potentially generate early warnings about unknown ADRs. Following on from this work, various lexicon-based approaches for ADR detection from social media have been proposed [26–30,21,31]. However, since they have to rely on the development of lexicons, these works are generally limited in the number of drugs studied or the number of target ADRs (*e.g.*, [26] focuses on 4 drugs, [27] on 10 drugs and 5 ADRs, [28] on breast cancer associated ADRs only, and [29] on 5 drugs). Our intent at this point is to be able to detect user posts mentioning ADRs without taking into account any associated drug(s).

To address some of the limitations of the lexicon-based approach, Nikfarjam and Gonzalez [32] proposed a method for capturing the underlying syntactic and semantic patterns from social media posts. Machine learning based approaches have also been applied for ADR relation extraction [10,30] and for the classification of drugs into categories [11], including our pilot studies in this area [7,33]. However, there is still no research that attempts to perform deep linguistic analyses of social media text to generate rich features. With the emergence of available annotated data, there has been some research very similar to ours (*i.e.*, on the binary classification of ADR assertive text). For example, Gurulingappa et al. [6] used the publicly available ADE corpus [34] to perform binary classification using various classifiers, obtaining a maximum *F*-score of 0.77 for the ADR class. Since this data set is publicly available, this provides us with the opportunity of evaluating our approach on this data and against this benchmark system.

Very recently, research has focused on combining social media data (*e.g.*, from Twitter, DailyStrength and others) with other sources of information for ADR detection. For example, Yeleswarapu et al. [21] combine user posted comments from social media with data from structured databases and MEDLINE abstracts, and show that these sources can be utilized for ADR detection. Freifeld et al. [31] present an approach for using user posts from Twitter ADR detection using a dictionary based approach. However, the approach relied on the manual categorization of ADR containing tweets. These very recent works utilizing social media data further motivate the need for effective automatic classification approaches that are capable of filtering social media content for ADR detection techniques.

A major obstacle to applying supervised learning approaches for social media-based ADR detection tasks has been the lack of publicly available annotated data. Other than our recent prior work [7], we only found two data sets that are publicly available [28,35]. However, both these data sets are very small. [28] contains 247 annotated instances containing ADR mentions, and [35] contains 400 annotated tweets in Spanish. Thus, there is a strong motivation for the large scale annotation of data, such as the one we describe in the following section.

3. Data

We use three data sets for the research tasks described in this paper. Two of these were developed in-house, while the third data set is publicly available and has been sourced form medical case reports. All three data sets consist of text segments (*e.g.*, sentences, groups of sentences, and micro-blogs) that have been manually annotated for the presence or absence of ADRs. We briefly discuss all three data sets in this section, including the details of the creation of our in-house data sets.

3.1. Twitter

Our first data set has been sourced from the social networking site Twitter.[3] As mentioned earlier, Twitter is an extremely popular *micro-blogging* site with over 645,000,000 users and growing rapidly. The corpus was created during the first phase of annotations of a large study on ADR detection from social media that is currently in progress. We have made part of this growing corpus publicly available for research purposes.[4]

The first step in our data collection process involved the identification of a set of drugs to study, followed by the collection of user comments associated with each drug name. To maximize our ability to find relevant comments, we focused on two criteria: (i) drugs prescribed for chronic diseases and conditions that we might expect to be commonly commented upon and (ii) prevalence of drug use. For the first criterion, we selected drugs used to treat chronic conditions such as type 2 diabetes mellitus, coronary vascular disease, hypertension, asthma, chronic obstructive pulmonary disease, osteoporosis, Alzheimer's disease, overactive bladder, and nicotine addiction. To select medications that

[3] https://twitter.com/.

[4] diego.asu.edu/downloads/.

have a relatively high prevalence of use and thus exposure, we selected drugs from the IMS Health's Top 100 drugs by volume for the year 2013. The final drug list was prepared by a pharmacology expert, and for the data set used for the experiments described in this paper, a total of 74 drugs were used.

The tweets associated with the data were collected using the generic and brand names of the drugs, and also their possible phonetic misspellings [36], 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, which consisted of 10,822 instances. The data was annotated by two domain experts under the guidance of a pharmacology expert. Each tweet is annotated for the presence of ADRs, spans of ADR indications, and beneficial effects. For the research described here, we use the annotations associated with the presence of ADRs. Following the annotation of the full set, the disagreements were resolved by the pharmacology expert. In addition, a randomly chosen subset of the data (10%; 1082 tweets) was annotated by the pharmacology expert for the measurement of Inter Annotator Agreement (IAA). We used Cohen's Kappa [37] to compute IAA which is given by the following equation:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \qquad (1)$$

where $Pr(a)$ is the relative observed agreement among annotators, $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. We computed κ for all three pairs of agreements, and obtained an average of 0.71, which can be considered as *significant agreement* [38]. For the two annotators, $\kappa = 0.69$.[5] The final data set is highly imbalanced, as one would expect, with 1239 (11.4%) tweets containing ADR mentions and 9583 (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 pilot study paper [7]. For the rest of this paper, we will refer to this data set as TW.

3.2. DailyStrength

Our second data set has been sourced from the online health community DailyStrength. DailyStrength has specific review pages for each drug and it allows patients to share their personal knowledge and experiences regarding diseases, treatments, and also join various disease-related support groups. It serves as a resource for patients to connect with others who have similar conditions, and attracts more than 300,000 monthly visitors.[6]

At the time of writing of this paper, the data set consisted of a total of 10,617 annotated comments containing a total of 56 drugs. The drugs were chosen the same way as the TW set, by the pharmacology expert, and all posts associated with the drugs that were available up to March, 2014 were crawled from the social network. Since each drug has a specific review page in DailyStrength, all the comments associated with a drug were collected from the appropriate review pages without requiring any filtering. From over 25,000 posts, 10,617 were randomly chosen for binary annotation. Unlike the TW data, the posts in this data set contain more structure, are longer, and often consist of multiple sentences. The annotation was performed at the comment level in an identical fashion as the TW set. The agreement among the annotators was similar to the TW set, with $\kappa = 0.67$. This data set also exhibits significant imbalance, although not as much as the TW data set. It contains 2513 (23.7%) instances containing ADR mentions and 8104 (76.3%) instances not containing any ADR mentions. Further details about this data set can be found in our pilot study [33]. For the rest of this paper, we will refer to this data set as DS.

3.3. ADE corpus

The third data set we use is the publicly available ADE corpus [10]. This data set consists of double-annotated sentences, indicating the presence/absence of ADRs, obtained from medical case reports. Although the source of this text is not social media, we include this data set in our research for three reasons: (i) to compare the performance of our approach with state-of-the art approaches for this task, (ii) to investigate how the performance of classification algorithms vary between data sets of different origins, and (iii) to explore the possibility of utilizing this data for training learners that may be applied to classifying data from other sources. The latest version of this data set contains a total of 23,516 instances with 6821 (29.0%) containing ADR mentions and 16,695 (71.0%) not containing any ADR mentions. For the rest of this paper, we will refer to this data set as ADE. Detailed description about this data set is provided in [34]. Table 1 compares the sizes and distributions of the three data sets.[7]

4. Classification methods

In this section we discuss our automatic classification experiments in detail. The section is divided into two broad subsections. In the first subsection, we provide details of our binary classification approach for each of the three data sets. We provide detailed descriptions of the features that we extract from the data using NLP methods, explaining how we expand the feature space for the instances by generating various semantic information. In the second subsection, we discuss our approaches for multi-corpus training of data, and describe the various experiments that we perform to assess the effectiveness of multi-corpus training.

4.1. Binary classification

The first problem we address is the binary classification of text segments into the ADR or non-ADR categories. We perform the classification using three supervised classification approaches: Naïve Bayes (NB), Support Vector Machines (SVM) and Maximum Entropy (ME). SVMs are a natural choice for a text classification problem such as this as they have been shown to perform particularly well for such tasks due to their capability to deal with high dimensional feature spaces, dense concept vectors, and sparse instance vectors [39]. We use the ME classifier primarily because it has been shown to perform extremely well particularly for this classification problem in the past [6].

For the classification experiments, we divide each of the three data sets into two parts: 80% for training and 20% for evaluation. These 80/20 splits are stratified, meaning that the proportions of the ADR and non-ADR instances are the same as in the full data sets. We now provide a detailed description of the text preprocessing approaches and the feature sets we employ for classification.

4.1.1. Preprocessing

We perform standard preprocessing such as tokenization, lowercasing and stemming of all the terms using the Porter stemmer[8]

[5] Note that when computing the κ for the three pairs, the set of 1082 instances that have been annotated by all three team members are used. For the κ value given for the agreement between the two annotators, the whole set (10,822) instances is used.

[6] Based on: http://social-networking.findthebest.com/l/33/DailyStrength. Accessed on April 22, 2014.

[7] Our collection and annotation of the data for the TW and DS sets are ongoing, and these data sets will be larger once more phases of annotation are completed.

[8] We use the implementation provided by the NLTK toolkit http://www.nltk.org/.

Table 1
Distribution of ADR and non-ADR instances for the three data sets.

Data set	ADR instances	Non-ADR instances
TW	1239	9583
DS	2513	8104
ADE	6821	16,695

[40]. Our preliminary investigations suggested that stop words can play a positive effect on classifier performances by their presence in word 2-grams and 3-grams; so we do not remove stop words from the texts.

We parse the texts using two different parsers/taggers. Since two of our data sets come from social media, we apply the *Twitter Part-Of-Speech (POS) Tagger*[9] [41] to parse the texts and identify POS tags. However, other than identifying the POS tags for each term, this parser does not provide any additional information. Therefore, we also parse the texts using the Stanford parser[10] [42]. The Stanford parser is a lexicalized probabilistic parser which provides various information such as the syntactic structure of text segments, dependencies and POS tags. It can also be used as an accurate, unlexicalized stochastic context-free grammar parser.

In addition to parsing, we identify all the Unified Medical Language System (UMLS) medical semantic types and concept IDs (CUIs) from the text using the MetaMap[11] toolbox [43]. The UMLS is a compendium of many controlled medical vocabularies, and provides a mapping structure among the different vocabularies. The UMLS semantic types represent broad categories of medical concepts such as *disease or syndrome* and *pharmacological substance*. The CUIs provide more fine grained categorization of medical concepts so that different lexical representations of the same concept can be identified (*e.g.*, *hypertension* and *high blood pressure*). Besides using MetaMap to identify the semantic types and CUIs, we also use MetaMap's negation detection system to identify concepts that have been negated in the texts. MetaMap uses the NegEx[12] [44] tool identify and tag negations.

Finally, we perform some basic preprocessing that are specifically applicable to social media text. In particular, we remove references to Twitter user names by removing terms starting with the @ character, remove the # character from *hashtags*, and remove hyperlinks to other web pages.

4.1.2. N-grams

Our first feature set consists of word *n*-grams of the comments. A word *n*-gram is a sequence of contiguous *n* words in a text segment. This feature enables us to represent a document using the union of its terms. We use 1-, 2-, and 3-grams as features.

4.1.3. UMLS semantic types and concept IDs

The UMLS semantic types and CUIs present useful medical concepts from the text segments. As such we use them as features. For this feature set, we compute the *Tf.Idf* values [45] for the semantic types and CUIs. *Tf.Idf*—abbreviation for term frequency–inverse document frequency—is a popular measure in NLP and reflects how important a term is to a text segment. Terms that are unique to specific text segments, get higher scores, while terms that are common throughout the corpus are assigned lower scores by this measure.

4.1.4. Syn-set expansion

It has been shown in past research that certain terms, because of their prior polarities, play important roles in determining the polarities of sentences [46]. However, polarity indicating features have not been utilized for ADR detection in the past, and the utilization of such information for ADR classification is novel. Certain adjectives, and sometimes nouns and verbs, or their synonyms, are almost invariably associated with positive or non-positive polarities. For each adjective, noun or verb in a sentence, we use WordNet[13] to identify the synonyms of that term and add the synonymous terms, attached with the SYN tag, as features. Similar to the previous feature set, we use the *Tf.Idf* measure for each derived synonym. Terms that are flagged by MetaMap to be negated, are not included when generating the synonyms.

4.1.5. Change phrases

This polarity indicating feature set was proposed by Niu et al. [47], and the intuition behind this feature set is that whether a sentence represents a positive information or a negative information can often be signaled by how a change happens: if a bad thing (*e.g.*, headache) was reduced, then it is a positive outcome; if a bad thing was increased, then the outcome is negative. This feature set attempts to capture cases when a good/bad thing is increased/decreased. We first collected the four groups of good, bad, more, and less words used by Sarker et al. [46]. This feature set has four features: MORE-GOOD, MORE-BAD, LESS-GOOD, and LESS-BAD. To extract the first feature, we applied the same approach as [47]: a window of four words on each side of a MORE-word in a sentence was observed. If a GOOD-word occurs in this window, then the feature MORE-GOOD is activated. The other three features were activated in a similar way.

4.1.6. ADR lexicon matches

This is a novel feature for ADR classification, and the key idea behind this feature is to incorporate domain-specific knowledge to the classification process by performing ADR lexicon matches. Although lexicon-based approaches have several limitations for ADR extraction, we suspect that lexicon match information is likely to be useful for ADR classification. We use a lexicon built in-house to detect ADRs from the text segments. Our lexicon was derived from the lexicon used by Leaman et al. [8], which includes terms and concepts from four resources. These include the COSTART[14] vocabulary created by the FDA for post-market surveillance of ADRs, and contains 3787 concepts; the SIDER[15] side effect resource which contained 888 drugs linked with 1450 ADR terms at the time of our lexicon creation; and the Canada Drug Adverse Reaction Database (MedEffect[16]), which contained associations between 10,192 drugs and 3279 adverse reactions at the time of our lexicon creation. These three resources contain concept names and UMLS CUIs, and the lexicon was manually reduced by grouping terms with similar meanings (*e.g.*, *appetite exaggerated* and *appetite increased*). We added additional terms from SIDER II [48] and the Consumer Health Vocabulary Initiative[17] [49]—which include colloquialisms. Our final lexicon contains a total of 16,183 ADR mentions: including both standard and non-standard terms.

We derive two features using the ADR lexicon. The first feature is a binary feature indicating the presence/absence of ADR mentions, as identified by string matching using the lexicon. The

[9] http://www.ark.cs.cmu.edu/TweetNLP/. Accessed on October 13, 2014.

[10] http://nlp.stanford.edu/software/lex-parser.shtml. Accessed on October 13, 2014.

[11] http://metamap.nlm.nih.gov/. Accessed on October 13, 2014.

[12] Available from: https://code.google.com/p/negex/. Accessed on October 13, 2014.

[13] http://wordnet.princeton.edu/. Accessed on October 13, 2014.

[14] http://www.nlm.nih.gov/research/umls/sourcereleaseddocs/current/CST/. Accessed on October 13, 2014.

[15] http://sideeffects.embl.de/. Accessed on October 13, 2014.

[16] http://www.hc-sc.gc.ca/dhp-mps/medeff/index-eng.php. Accessed on October 13, 2014.

[17] http://www.consumerhealthvocab.org/. Accessed on October 13, 2014.

second feature is a numeric feature computed by counting the number of ADR mentions in a text segment and dividing it by the number of words in the text segment.

4.1.7. Sentiword scores

Our inspection of the data suggests that comments associated with ADRs generally present negative sentiment. Sentiment analysis is a field of research which focuses primarily on distinguishing between positive and negative sentiment from lexical data [50]. For this feature, we incorporate a score that attempts to represent the general sentiment of a comment. Each word-POS pair in a comment is assigned a score and the overall score assigned to the comment is equal to the sum of all the individual term-POS sentiment scores divided by the length of the sentence in words. Terms identified by MetaMap to be negated are not included and for term-POS pairs with multiple senses, the score for the most common sense is chosen. To obtain a score for each term, we use the lexicon proposed by Guerini et al. [51]. The lexicon contains approximately 155,000 English words associated with a sentiment score between −1 and 1. The overall score a sentence receives is therefore a floating point number with the range [−1:1].

4.1.8. Topic-based feature

Our last novel feature is based on topic modeling [52], an approach that attempts to discover abstract topics that occur in collections of texts. Our intuition is that ADR assertive text segments are likely to exhibit specific abstract topics, which may be captured and utilized for text classification. We use the Mallet[18] tool to generate topics from each of the three training set ADR instances. The tool generates keywords associated with each abstract topic, and a score indicating the relevance of the topic to the full set of texts. We use this information to generate two features from each instance: (i) the topic terms that appear in the instance and (ii) the sums of all the relevance scores of the terms in each instance.

4.1.9. Other features

In addition to the features already mentioned, we use some simple features, which are as follows:

(i) Length: the lengths of the text segments in words.
(iii) Presence of comparatives and superlatives. These are binary features and these items are identified from the Stanford parses of the text segments.
(iii) Presence of modals.

4.1.10. Classifier settings

For the NB and ME classifiers, we use their default implementations in the machine learning tool Weka.[19] For the SVM classifier, we use the LibSVM implementation[20] [53]. The performance of SVMs can vary significantly based on the kernel and specific parameter values. For our work, based on some preliminary experimentation on the training set, we use the RBF kernel. We compute optimal values for the *cost* and γ parameters via grid-search and 10-fold cross validation over the training sets. To address the problem of data imbalance, we utilized the weighted SVM feature of the LibSVM library. We gave a higher weight (w_1) to the smaller class (c_1), and the value for the weight was computed from the equation: $w_1 \times size(c_1) = size(c_2)$. Finally, we performed scaling of the feature vectors before the classification process so that all the feature values were in the range [0:1].

4.2. Multi-corpus training and classification

Following the separate binary classification of each of the three data sets, we attempt to combine data from different corpora to assess how that affects classification accuracies. Due to the largely imbalanced data set, for the TW and DS data sets particularly, there are not many ADR instances. It is likely that the availability of more training data would improve classification performance for these data sets. However, preparation and annotation of data is an expensive process, and especially to increase the number of instances for the minority class, large volumes of data would require annotation (*e.g.*, approximately 8000 tweets need to be annotated to increase the number of ADR class instances by 1000). Therefore, the combination of multiple corpora for improving classification performance is an interesting possibility, and its investigation is crucial. In this subsection, we describe how we combine training instances from distinct data sets to investigate how they influence classification accuracies.

Since we generate a number of portable features from the texts, we expect that if the data sets are compatible, they will support cross training. For each of the three data sets, we perform two sets of experiments, each set with one of the other two data sets. For example, for the TW data set, we combine the ADR instances from the DS data set for one set of experiments, and the ADR instances from the ADE data set for the second set of experiments. Following these experiments, we perform another pair of experiments for each data set, this time combining all the training instances of the two other data sets. These two sets of experiments enable us to determine how the classification performance over the ADR class changes, if at all, when (i) only the ADR instances of another data set are added to the training data and (ii) when all the instances of another data set are added to the training data.

For the final set of experiments, we combine the instances from all three data sets and attempt to classify the test set instances for each of the data sets. One major issue encountered when performing these sets of experiments is that adding only the ADR instances makes the number of ADR instances in the training set larger than the number of non-ADR instances, and this makes the balance of the training set significantly different compared to the test set. Note that in all of the *paired data* experiments described in the previous subsection, the number of ADR instances was always lower than the number of non-ADR instances, but this changes if we add all the ADR instances from the three data sets. Therefore, for this set of experiments, we always use both classes of instances.

5. Evaluation, results and discussion

In this section, we provide details of the classification results of our various experiments. Like the previous section, this section is divided into two broad parts—the first presenting the results of our classification task over individual data sets and the second providing results of the multi-corpus classification experiments.

5.1. Binary classification results

Among the three classifiers, SVMs perform significantly better than the other two with ADR *F*-scores of 0.812, 0.538 and 0.678 for the ADE, TW and DS data sets, respectively. For the ADE data set, our SVM classifier performs significantly better than the previously reported highest *F*-score of 0.77 by Gurulingappa et al. [6][21]. Similarly, for the DS data set, our SVM classifier outperforms our

[18] http://mallet.cs.umass.edu/index.php. Accessed on October 13, 2014.
[19] http://www.cs.waikato.ac.nz/ml/weka/. Accessed on October 13, 2014.
[20] http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Accessed on October 13, 2014.
[21] Statistical significance of the *F*-score measured using the technique proposed by [54].

Table 2
Paired classification performances over the three data sets when only ADR instances from a different corpus are added. ADR *F*-scores, non-ADR *F*-scores, Accuracies and 95% Confidence Intervals (CI) for each of the train-test set combinations are shown.

Test data	Training data	ADR *F*-score	non-ADR *F*-score	Accuracy (%)	95% CI
ADE	ADE	0.812	0.914	88.2	87.3–89.1
	$\text{ADE} + \text{DS}_{ADR}$	0.802	0.912	87.8	86.8–88.7
	$\text{ADE} + \text{TW}_{ADR}$	0.802	0.909	87.5	86.5–88.4
TW	TW	0.538	0.919	86.2	84.7–87.6
	$\text{TW} + \text{ADE}_{ADR}$	0.549	0.946*	90.3*	89.0–91.5
	$\text{TW} + \text{DS}_{ADR}$	0.565*	0.939*	89.3*	87.9–90.5
DS	DS	0.678	0.890	83.8	82.2–85.0
	$\text{DS} + \text{ADE}_{ADR}$	0.682	0.886	83.2	82.7–85.8
	$\text{DS} + \text{TW}_{ADR}$	0.695*	0.897	84.6	82.8–86.0

* Statistically significant improvement in performance over the highest score achieved in the single corpus binary classification task.

previously reported ADR *F*-score of 0.652 [33]; and for the TW data set, our previously reported *F*-score of 0.529 [7].[22]

Our experimental results clearly suggest that the rich set of features we generate via the use of NLP, and the tuning of the SVM parameters, provide significant improvements over past approaches. An interesting and promising discovery from these experiments is that text classification techniques from the research areas of sentiment analysis, subjectivity analysis, topic modeling, and polarity classification can be used to improve automatic ADR detection techniques from internet-based media. However, while the ADR class *F*-score for the ADE data set is relatively high (0.812), the *F*-score values for the other two data sets are low (particularly for the TW data set). One obvious reason for this is the low number of training instances that we have available for the ADR class in the TW data set (and also the DS data set). To increase the size of the ADR instances in the data sets, more annotation is required, which, as mentioned earlier, is expensive. We, therefore, investigated the possibility of solving this imbalance problem, at least partially, by performing multi-corpus training.

5.2. Multi-corpus classification results

Tables 2 and 3 show the classifier performances for the paired classification experiments. In each of the two tables, the best performance obtained using a single data set for training is shown for comparison. A number of interesting and important results can be observed from the tables. First of all, the classification performances over the ADR instances significantly benefit from multi-corpus training for the DS and TW data sets, particularly when combined with each other. Adding the ADE training instances to either of the other two data sets also improves performance, although in most cases the improvements are not statistically significant.[23] For the ADE data set, there were no significant differences in performance compared to the single corpus binary classification experiments. This can be due to two reasons. Firstly, as explained earlier in the paper, the ADE data set does not contain data from social media. Therefore, the contents of the text segments are likely to be different compared to the DS and TW data sets. Thus, adding the data from the other two sets does not improve performance. Secondly, the ADE data set already contains approximately twice the amount of instances compared to the other two data sets. As a result, it is possible that a larger amount of data is required to further improve the classification performance over the ADE data set. Considering the fact that combining DS and TW sets results in better performances over these two data sets relative to the performance when ADE instances are added to the training, the first of the two explanations seems more likely to be correct.

When all three data sets are combined for training, we do not observe any further improvements compared to the paired classification experiments, as shown in Table 4. This observation is in line with the paired classification experiments which show that adding training data from the ADE data set does not significantly improve performances for the TW and DS classes. This suggests that although multi-corpus training can be applied to improve performance, such approaches are likely to work when similar data sets are combined, and not when the data sets are dissimilar.

5.3. Impact of multi-corpus training

We wanted to compare the *projected* result for a specific data set size and the actual result for that specific size of training set obtained by combining data from another data set. For example, when training using the TW + DS data sets and testing on the TW test set, we added all the DS training instances (8494) to the TW instances. We are interested in predicting what the ADR *F*-score on the TW test set would be if an equal number of TW instances were added for training, and compare that value to the actual performance when the DS instances are used for training instead. This will give us an estimate of the performance gap between training using data from the same corpus data and using data from another corpus. To do this analysis, we first generated training set size *vs.* ADR *F*-score graphs for the data sets. We then fit trendlines to the ADR *F*-scores. We found logarithmic trendlines to have the best fit, and found that the ADE data set has the best fitting trendline and the TW data set to have the least fitting trendline. Importantly, using the equations for the trendlines, we were able to estimate the projected performances and compare them with the actual performances using data from a different corpus. Because we found the TW and DS data sets to be *compatible* for multi-corpus training, we only performed this analysis for these two data sets. Fig. 3 summarizes the results of this comparison. The figure shows that for the DS set, the projected score is very close to the actual *F*-score obtained when combined with the TW data set. However, for the TW set, it can be seen that the difference is greater. In fact, for the TW set, because of the very low values towards the left of the figure, the trendline clearly over estimates the projected score. The R^2 value also shows that the trendline for the DS set has much better fit than the TW set. These experiments demonstrate the usefulness of multi-corpus training.

[22] The mentioned paper performs classification via under sampling, which yields higher ADR *F*-scores at the expense of overall accuracy. Furthermore, in the undersampling experiments in the mentioned paper, the test set is also artificially balanced (as the intent of the paper was to show the utility of the TW corpus and not to evaluate classification performances on real-life data), and thus, some of the experimental results do not represent performances over real-life data. As such, we compare against the 70–30 ratio discussed in that paper, which most closely represents real life data.

[23] Obtaining statistically significant improvements is quite challenging because of the already good performances of the classifiers in the single corpus binary classification tasks.

Table 3
Paired classification performances (all instances) over the three data sets. ADR *F*-scores, non-ADR *F*-scores, Accuracies and 95% Confidence Intervals (CI) for each of the train-test set combinations are shown.

Test data	Training data	ADR *F*-score	non-ADR *F*-score	Accuracy (%)	95% CI
ADE	ADE	0.812	0.914	88.2	87.3–89.1
	ADE + DS$_{ALL}$	0.789	0.904	86.9	85.9–87.8
	ADE + TW$_{ALL}$	0.800	0.912	87.7	86.8–88.7
TW	TW	0.538	0.919	86.2	84.7–87.6
	TW + ADE$_{ALL}$	0.545	0.941	88.6	87.2–89.7
	TW + DS$_{ALL}$	0.597*	0.943	90.1	88.7–91.3
DS	DS	0.678	0.890	83.8	82.2–85.0
	DS + ADE$_{ALL}$	0.674	0.891	83.5	81.6–84.8
	DS + TW$_{ALL}$	0.704*	0.899	85.0	83.3–86.5

* Statistically significant improvement in performance over the highest score achieved in the binary classification task.

Table 4
Classification performance when all the available training data is used to perform classification on the test sets of the three data sets.

Test data	Training data	ADR *F*-score	non-ADR *F*-score	Accuracy (%)	95% CI
ADE	ALL three	0.799	0.913	87.8	86.8–88.7
TW	ALL three	0.564	0.934	88.5	87.1–89.8
DS	ALL three	0.686	0.887	83.4	81.7–84.9

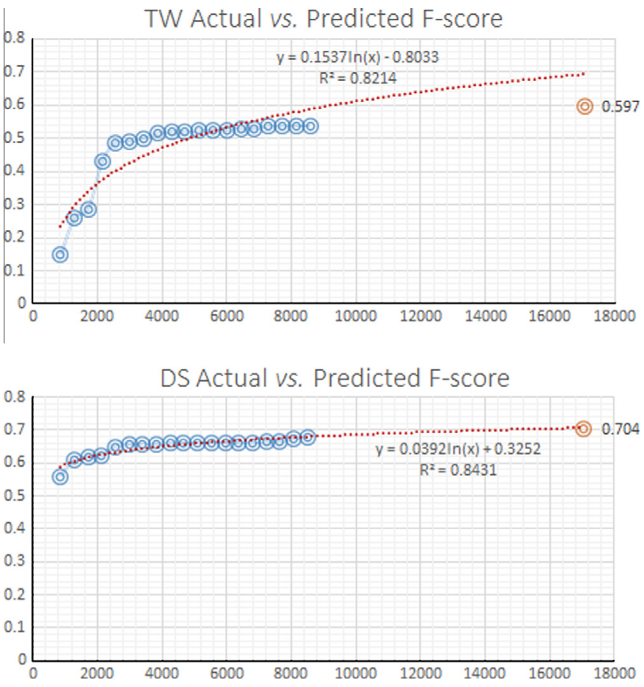

Fig. 3. Comparison of actual *vs.* projected accuracies for the DS and TW data sets.

5.3.1. Contribution of features

We performed several *leave-one-out* classifications to investigate the contribution of each of the features on classification performances. Table 5 presents the results of these experiments. It can be observed from the table that all the features tend to improve ADR *F*-scores, and the *F*-score drops when they are removed from training. The only case when the ADR *F*-score improved is for the *syn-set expansion* feature on the TW data set; this, however, is a very small improvement, and is not statistically significant. In terms of impact, *n-grams* tend to be very useful, as expected, and there are significant drops in the ADR *F*-scores when they are removed. Similarly, the *UMLS semantic types and CUIs* clearly have a crucial impact in classification performance and large drops are seen when they are removed during training. The *syn-set expansion* and *topic model* features appear to be particularly useful for the ADE data set. For the *syn-sets*, this can be due to the fact that there are generally more synonyms associated with the terms in the ADE instances than the two other data sets or because WordNet is unable to identify synonyms for a number of colloquial or misspelt terms. For the *topic models*, it is perhaps due to the fact that the ADE corpus has a much larger amount of ADR information, resulting in the generation of more reliable topics. The *sentiword* feature, in contrast, appears to be more useful for the social media based data sets compared to the ADE corpus. This may suggest that users tend to use terms with negative prior polarities when discussing ADRs. This illustrates the benefits of utilizing sentiment indicating features when mining ADRs from social media. The

Table 5
Leave-one-out classification scores over the three data sets showing how accuracies and ADR *F*-scores are affected as one feature is removed from the set.

Features	TW		DS		ADE	
	Accuracy	ADR *F*-score	Accuracy	ADR *F*-score	Accuracy	ADR *F*-score
All	86.2	0.538	83.6	0.678	88.2	0.812
N-grams	80.7	0.424	82.6	0.654	85.9	0.775
UMLS STs and CUIs	85.7	0.505	82.8	0.652	81.9	0.711
Syn-set expansions	86.1	0.545	84.0	0.669	87.9	0.778
Change phrases	87.1	0.521	83.9	0.665	88.0	0.803
ADR lexicon match	86.1	0.492	83.5	0.663	86.1	0.780
Sentiword score	86.2	0.530	82.8	0.659	88.3	0.805
Topics	86.1	0.535	83.7	0.670	87.6	0.801
Other features	86.9	0.534	83.6	0.677	88.1	0.809

Table 6
Examples illustrating common reasons behind the misclassification of tweets.

Post	Classified as	Issues
geez this vyvanse makes me talk a mile a minute haha	non-ADR	ambiguous, short
ok vyvanse, it's 4:30am. at this point ima just throw you a big **** you	non-ADR	ambiguous reference to ADR (sleeplessness)
this cipro is totally "killing" my tummy ..hiks..	non-ADR	non-standard description of ADR
i don't know if vyvanse was a good idea #nosleep	non-ADR	ambiguous, non-standard description of ADR
#restlesslegs #quetiapine	non-ADR	short, non-standard terms
had a dream my eyes got burnt out of my head, thank you trazodone	non-ADR	sarcasm/contradictory statement
popped a vyvanse, im tweaked. whoop	non-ADR	ambiguous, short, non-standard terms
fluexetine and quet zombified me..ah, the meds merrygorount #bipolar	non-ADR	spelling mistakes, non-standard terms
wtf Humira, 'clears your skin but may cause u cancer or heart failure smh'...	ADR	generic statement about ADR, no reference to personal experience
shorter xarelto advertisements: if you take it you may "bleed to death"	ADR	general ADR statement (reference to advertisement)

change phrases and other features tend to have relatively low impacts on all three data sets. Finally, the DS data set tends to show the lowest variance between the experiments and the TW set tends to show the highest. Importantly, it is observed that the best score is obtained by the combination of all the features and not by individual features of subsets of the features used. This supports our initial hypothesis that the generation of a rich set of features along multiple dimensions improves the performance of learning-based approaches. Note that due to the inter-dependence of some of the features (*e.g.*, negation with sentiment score and syn-sets), we only assess the impacts of the different features using the leave-one-out scores rather than individual feature scores. Also, some of the features are only generated for a limited number of instances (*e.g.*, change phrases in the TW set), meaning that the benefits of these features can only be observed when combined with other features and not when applied independently.

5.3.2. Error analysis

We conclude this section with an analysis of the causes of classification errors, and a discussion possible techniques that can be applied/incorporated in the future to address these problems.[24] The common causes for misclassifications are as follows:

- *Non-standard terms/use of English*—Many ADR descriptions are described using non-standard terminology. These are often very unique to specific posts and are not repeated. As a result, the classifiers often misclassify these cases.
- *Short posts*—Some posts are very short and consist of very few terms. This makes the generation of a rich set of features associated with the posts impossible, thus, resulting in misclassifications. Due to the large number of non-ADR instances, such posts get classified as such, even when they mention ADRs.
- *Large proportion of spelling errors*—Often posts consist of large numbers of misspellings. As the misspelt terms do not contribute to the sentiment/polarity/lexicon/topic scores, these posts are often misclassified (to the non-ADR class).
- *Generic statements about ADRs, not personal experience*—Often users mention multiple ADRs in their posts, but the statements are generic and do not represent personal experiences. This is often the case when users quote lines from the disclaimers in the advertisements for the specific drugs. This results in misclassification of non-ADR comments as ADRs.
- *Mixed sentiments/ambiguous statements*—Often the user posts are very ambiguous and do not clearly represent their intents. This is often also the result of sarcastic statements. In such statements, positive feelings are often expressed regarding ADRs. This generally results in ADR comments being classified as non-ADR.

A large number of user posts exhibit multiple of the above issues and prove to be very difficult to classify. Table 6 presents ten sample user posts, describing the issues they exhibit.

5.4. Summary of results and discussions

We conclude this section with the following brief list to summarize our research findings.

- By using NLP techniques to extract novel features from text, the performance of machine learning algorithms attempting to automatically detect ADR assertive sentences may be significantly improved.
- Combining lexical features from well-established research areas such as sentiment analysis/polarity classification can improve the automatic classification of ADR mentions from social media text. This is perhaps because users generally express sentiments when posting about drug associated events in social media, and the sentiments generally correlate strongly with the reactions associated with the drugs they are taking.
- The two previously unused features, based on ADR lexicons and topic models, boost ADR classification performances, the latter being more useful when a large amount of text is available for the generation of topics.
- Multi-corpus training can provide significant improvements in classification accuracies if the corpora used are compatible. Considering the numerous social networks that are available today, this provides the opportunity to combine data from various sources to train machine learning algorithms. This also means that significantly lower amounts of time can be spent on annotation of data, and instead, annotated data from distinct sources can be combined to perform ADR detection tasks from social media. Systems trained on data from multiple social media sources are likely to be more portable than those trained from a single corpus.

6. Conclusions and future work

In this paper we focused on the problem of automatic text classification of sentences to detect ADR mentions. In particular, we attempted to investigate approaches by which NLP and machine learning techniques can be applied to improve the automatic classification of social media text containing medical information. Our particular focus was social media data, but we also included a publicly available data source from outside the social media domain for scientific comparison. We discussed the collection and preparation of two data sets that we have prepared in-house, one of which has been made publicly available. We utilized NLP techniques to extract useful features from the three data sets and showed that the careful selection of features can significantly improve automatic classification accuracies. The features chosen were indicative of various properties of the texts including their polarities, sentiments and topics. Our experimental results strongly suggest

[24] This analysis only refers to the social media component of our research (*i.e.*, the TW and DS sets).

that the combining of features from well established text classification research areas, such as sentiment analysis, can improve classification accuracies. As such, there is a strong motivation for further exploration of similar well established research areas to identify features that may be applied to automatic ADR detection and monitoring. Furthermore, the topic model feature that we introduced is likely to benefit text classification tasks in general, and should be explored in future research.

We explored the possibility of combining annotated data from different corpora in an attempt to improve classification accuracies. Our experiments showed that significant improvements in classification accuracies can be achieved by combining the DS and TW data sets, and slight improvements can be achieved over the DS and TW sets when combined with the ADE training data. This suggests that if data sets are compatible (*i.e.*, they contain similar types of text), they can be combined for training. Although various properties of the DS and TW data sets are very different (*e.g.*, post lengths, contents, *n*-gram models, and so on), the use of features, such as indication of sentiments, topics, ADRs, and polarities, maybe used to reflect compatible properties of the texts. This leads to improved performances when cross-corpus training is performed, and similar techniques can be used in the future for text classification problems attempting to utilize multiple data sets. The empirical validation of this hypothesis may have crucial implications, particularly when working with social media data, where the volume of data is colossal, and numerous sources of information exist. This may also significantly reduce annotation time and expenses.

In the future, we would like to combine information from multiple corpora for the task of ADR relation extraction, which is the next step in our ADR monitoring pipeline. Considering that we achieved significantly better results than past research in the area, we expect that combining information from multiple sources for the later stages of our work will produce similar results. From an NLP and more general medical informatics perspective, we want to explore what NLP and feature extraction techniques from various text classification problems can be applied to text in the medical domain. It must be mentioned that the key to improving classification performance via combining corpora was the use of NLP on distinct data sets to extract features which represented similar information across the corpora. Thus, future research in this area may provide further break-through. Finally, there has been major progress in automatic text summarization research in the medical domain, such as the work presented in [55]. Considering the large volume of drug related information available in the various social networks, we will attempt to use summarization techniques to compress and identify critical information about drugs and utilize the summaries to assess drug safety.

7. Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank the NIH/NLM support for this Project (Award: 1R01LM011176-01).

References

[1] The importance of pharmacovigilance – safety monitoring of medicinal products. World Health Organization; 2002. <http://apps.who.int/medicinedocs/en/d/Js4893e/1.html>.

[2] Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012;91(3):1010–21.

[3] Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 1998; 279(15):1200–5.

[4] Ahmad SR. Adverse drug event monitoring at the food and drug administration – your report can make a difference. J Intern Med 2003;18(1):57–60.

[5] Xu R, Wang Q. Large-scale combining signals from both biomedical literature and FDA adverse event reporting system (FAERS) to improve post-marketing drug safety signal detection. BMC Bioinformatics 2014;15(17).

[6] Gurulingappa H, Fluck J, Hofmann-apitius M, Toldo L. Identification of adverse drug event assertive sentences in medical case reports. In: Proceedings of ECML PKDD 2-11 workshop on knowledge discovery in health care and medicine; 2011.

[7] Ginn R, Pimpalkhute P, Nikfarjam A, Patki A, O'Connor K, Sarker A, et al. 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.

[8] Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. 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. p. 117–25.

[9] Sharif H, Abbasi A, Zafar F, Zimbra D. Detecting adverse drug reactions using a sentiment classification framework. In: Proceedings of the sixth ASE international conference on social computing (SocialCom); 2014.

[10] Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug event from medical case reports. J Biomed Semantics 2012;3(15).

[11] Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. In: Proceedings of the American medical informatics association (AMIA) annual symposium; 2011. p. 217–26.

[12] Wadman M. News feature: strong medicine. Nat Med 2005;11:465–6.

[13] Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: proceed with caution. Pharmacoepidemiol Drug Saf 2007;16(4): 359–65.

[14] Bate A, Evans SJ. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf 2009;18(6):427–36.

[15] Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009; 16:328–37.

[16] Friedman C. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In: Proceedings of the 12th conference on artificial intelligence in medicine (AIME); 2009. p. 1–5.

[17] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010;160(Pt 1):739–43.

[18] Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2012; 20(3):413–9.

[19] Matthews EJ, Kruhlak NL, Benz DR, Aragone D, Merchant CA, Contrera JF. Identification of structure–activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q) SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. Regul Toxicol Pharmacol: RTP 2009;54(1):23–42.

[20] Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. J Am Med Inform Assoc 2011;18:668–74.

[21] Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R. A pipeline to extract drug-adverse event pairs from multiple data sources. BMC Med Inform Decis Mak 2014;14(13).

[22] Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. J Am Med Inform Assoc 2011;18(Suppl 1):73–80.

[23] Tuarob S, Tucker CS, Salathe M, Ram N. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. J Biomed Inform 2014;49:255–68.

[24] Edwards IR, Lindquist M. Social media and networks in pharmacovigilance: boon or bane? Drug Saf 2011;34(4):267–71.

[25] Davidson KP, Pennebaker JW, Dickerson SS. Who talks? The social psychology of illness support groups. Am Psychol Assoc 2000;55(2):205–17.

[26] Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. J Biomed Inform 2011;44:989–96.

[27] Yang CC, Yang H, Jiang L, Zhang M. Social media mining for drug safety signal detection. In: Proceedings of the 2012 international workshop on smart health and wellbeing; 2012. p. 33–44.

[28] Yates A, Goharian N. 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. p. 816–9.

[29] Jiang K, Zheng Y. Mining Twitter data for potential drug effects. Adv Data Min Appl 2013;8346:434–43.

[30] Liu X, Chen H. AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In: Proceedings of the 2013 international conference on smart health; 2013. p. 134–50.

[31] Freifeld CC, Brownstein JS, Menone CM, Bao W, Felice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Saf 2014;37(5):343–50.

[32] Nikfarjam A, Gonzalez G. 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. p. 1019–26.

[33] Patki A, Sarker A, Pimpalkhute P, Nikfarjam A, Ginn R, O'Connor K, et al. Mining adverse drug reaction signals from social media: going beyond extraction. In: Proceedings of BioLinkSig 2014; 2014.

[34] Gurulingappa H, Rajput AM, 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. J Biomed Inform 2012; 45:885–92.

[35] Segura-Bedmar I, Revert R, Martinez P. Detecting drugs and adverse events from Spanish health social media streams. In: Proceedings of the 5th international workshop on health text mining and information analysis (LOUHI); 2014. p. 668–74.

[36] Pimpalkhute P, Patki A, Gonzalez G. 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.

[37] Carletta J. Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 1996;22(2):249–54.

[38] Viera A, Garrett J. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):36–363.

[39] Joachims T. Text categorization with support vector machines: learning with many relevant features. Tech rep, Universitat Dortmund, Informatik LS8, Baroper Str 301, 44221 Dortmund, Germany; 1997.

[40] Porter MF. An algorithm for suffix stripping. Program 1980;14(3):130–7.

[41] Owoputi O, O'Connor B, Dyer C, Gimpel K, Schneider N, Smith NA. Improved part-of-speech tagging for online conversational text with word clusters. In: Proceedings of the NAACL-HLT, 2-13. p. 380–90.

[42] de Marneffe M-C, MacCartney B, Manning CD. Generating typed dependency parsers from phrase structure parses. In: Proceedings of the fifth international conference on language resources and evaluation; 2006. p. 449–54.

[43] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In: Proceedings of the American medical informatics association annual symposium; 2001. p. 17–21.

[44] Chapman WW, Hilert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. In: Proceedings of the 14th world congress on medical & health informatics (MEDINFO); 2013. p. 677–681.

[45] Rajaraman A, Ullman JD. Mining of massive datasets. Cambridge University Press; 2011 [Ch. 1] p. 1–7.

[46] Sarker A, Molla D, Paris C. Automatic prediction of evidence-based recommendations via sentence-level polarity classification. In: Proceedings of the international joint conference on natural language processing (IJCNLP); 2013. p. 712–8.

[47] Niu Y, Zhu X, Li J, Hirst G. Analysis of polarity information in medical text. In: Proceedings of the American medical informatics association (AMIA) annual symposium; 2005. p. 570–4.

[48] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6(343).

[49] Zeng-Treitler Q, Goryachev S, Tse T, Keselman A, Boxwala A. Estimating consumer familiarity with health terminology: a context-based approach. J Am Med Inform Assoc (JAMIA) 2008;15(3):349–56.

[50] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL conference on empirical methods in natural language processing (EMNLP); 2002. p. 79–86.

[51] Guerini M, Gatti L, Turchi M. Sentiment analysis: how to derive prior polarities from SentiWordNet. In: Proceedings of empirical methods in natural language processing (EMNLP); 2013. p. 1259–69.

[52] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.

[53] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2:27:1–27:27. Software. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

[54] Yeh A. More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th conference on computational linguistics, vol. 2; 2000. p. 947–53.

[55] Sarker A, Mollá D, Paris C. An approach for query-focused text summarisation for evidence based medicine. In: Proceedings of the 14th conference on artificial intelligence in medicine (AIME); 2013. p. 295–304.