Journal of Biomedical Informatics xxx (2014) xxx–xxx

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 xxxx

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/

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.

Please cite this article in press as: Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.11.002