Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use

Nestor Alvaro [a,b,*], Mike Conway [c], Son Doan [d], Christoph Lofi [e], John Overington [f], Nigel Collier [g,h]

[a] Department of Informatics, National Institute of Informatics, Japan
[b] The Graduate University for Advanced Studies, Japan
[c] Department of Biomedical Informatics, University of Utah, USA
[d] Medical Informatics, Kaiser Permanente Southern California, USA
[e] Institute for Information Systems, Technische Universität Braunschweig, Germany
[f] Stratified Medical, London, UK
[g] Department of Theoretical and Applied Linguistics, University of Cambridge, UK
[h] European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

ABSTRACT

Self-reported patient data has been shown to be a valuable knowledge source for post-market pharmacovigilance. In this paper we propose using the popular micro-blogging service Twitter to gather evidence about adverse drug reactions (ADRs) after firstly having identified micro-blog messages (also know as "*tweets*") that report first-hand experience. In order to achieve this goal we explore machine learning with data crowdsourced from laymen annotators. With the help of lay annotators recruited from CrowdFlower we manually annotated 1548 tweets containing keywords related to two kinds of drugs: SSRIs (eg. Paroxetine), and cognitive enhancers (eg. Ritalin). Our results show that inter-annotator agreement (Fleiss' kappa) for crowdsourcing ranks in moderate agreement with a pair of experienced annotators (Spearman's Rho = 0.471). We utilized the gold standard annotations from CrowdFlower for automatically training a range of supervised machine learning models to recognize first-hand experience. *F*-Score values are reported for 6 of these techniques with the Bayesian Generalized Linear Model being the best (*F*-Score = 0.64 and Informedness = 0.43) when combined with a selected set of features obtained by using information gain criteria.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The scale of serious and fatal adverse drug reactions (ADRs) has been a key focus of concern for public health systems, especially in the United States, since at least the turn of the century [1] with an estimated 100,000 deaths attributed to adverse drug reactions (ADRs) every year in US hospitals [2]. An ADR is defined as any noxious and unintended response to a medicinal product. We also understand an adverse drug event (ADE or AE) as any unfavourable and unintended sign, symptom, or disease temporally associated with the use of a medicinal product [3].

Given the limitations, and relatively small-scale of clinical trials for new drugs, post-market pharmacovigilance is vital. Traditional surveillance methods have focused on active clinician (or patient) reporting. The United States Food & Drug Administration's (FDA) Safety Information and Event Reporting Program (i.e. MedWatch) [4] collects reports from the pharmaceutical industry, but these typically undergo significant reporting delays and systematic under-reporting [5].

Social media has been shown to be a promising data source for pharmacovigilance data due to its real-time nature and utility in providing insights into off-label consumer habits [6,7]. Interest in social media as a signal source seems to be growing as can be seen by recent official announcements: On June 2014, the FDA presented its guidelines on how to use social media [8], and the Medicines and Healthcare products Regulatory Agency (MHRA) announced an application intended to report suspected ADRs, called WEB-RADR [9], on September 2014. EMA (European Medicines Agency) also published guidelines on good pharmacovigilance practices during 2013 [10] indicating that "*marketing authorisation holders should regularly screen internet or digital media*", and stating that web sites, web pages, blogs, vlogs, social networks, internet

forums, chat rooms, and health portals should be considered [11]. It seems clear that there is an increasing awareness of the potential for social media as a source of evidence. Our work here is focused on automatically identifying those Twitter messages that contain useful evidence for ADRs independently of whether these self reports comply with the guidelines or use the tools provided by the agencies mentioned above.

Twitter offers several potential benefits as a source for pharmacovigilance surveillance data. First, a significant fraction of the content is freely available via a public application programming interface (API). Second, the volume of data available is huge, and unmediated by gatekeepers, with approximately 500 million tweets sent per day in 2013 [12]. Third, Twitter content is *"real-time"*, allowing health researchers to potentially investigate and identify new ADE types faster than traditional methods such as physician reports. As such, we regard Twitter as an excellent testbed for our goal of identifying reports of ADRs among potential off-label drug users that may go under-reported by general practitioner visits [13] or undetected in clinical trials [14].

At least one potential unknown is the influence of population bias. Since Twitter users tend to have a particular demographic [15] this may influence the ability of the media to provide useful evidence for some classes of drugs, e.g. those drugs used primarily by paediatric and geriatric patients. In this study, we focus on two classes of drugs: Selective Serotonin Reuptake Inhibitors antidepressants (SSRIs) (e.g. fluoxetine, citalopram) and cognitive enhancers (e.g. modafinil, methylphenidate). SSRIs were selected due to public concerns regarding the risk of suicidal ideation in children and adolescents [16]. The cognitive enhancer drug category was chosen due to the wide spread off-label use of prescription drugs such as Ritalin and Adderall as study aids by university students [17].

A key difficulty in working with Twitter data, and social media data more generally, is distinguishing between first-hand experiences (*"I feel real groggy after taking <DRUG>"*), second-hand experiences (*"I've heard <DRUG> makes you real tired"*), and other kinds of information related to the drug, like news (*"Court found <DRUG> company liable"*) or advertising (*"Buy <DRUG> now!"*). In this paper we present a set of crowd-sourced Twitter annotations for SSRIs and cognitive enhancers, focusing on automatically identifying first-hand experiences. We show that annotations derived using the crowdsourcing service CrowdFlower are as reliable, in terms of inter-annotator agreement, as annotations derived from experienced annotators. Furthermore, we present a series of machine learning experiments based on these crowd-sourced annotations to show how first-person reports of ADRs can automatically be identified.

As a first stage in gathering data on ADRs, it is vital to identify first-hand drug usage experience. This is a challenging area for natural language processing (NLP) as social media messages contain a high proportion of ungrammatical constructions, out of vocabulary words, abbreviations and metaphoric usage. First-hand experience is defined as being where the person making the report has actually taken the drug. For example, *"<DRUG> is no joke have you up forever took it at 8 haven't been sleepy since #<HASHTAG> #<HASHTAG> #<HASHTAG>"*. On the other hand, a tweet like *"Think I'll just take some <DRUG> and get stuff done instead of sitting here like a worthless piece of shit."*, or *"New Years resolution. Be less boring by staying up past 8pm. #<HASHTAG> or <DRUG>"* would not be classified as first person as there is doubt as to whether the authors have taken the drug.

Previous studies [18] used a reduced set of drugs to compare the adverse events reported on social networks with the adverse events registered in official databases such as FAERS [19], but to the best of our knowledge no studies have explored the genre, i.e. the type of tweet, in which the users refer to the drugs.

2. Data selection

The drugs selected for our study were either cognitive enhancers, i.e. drugs that enhance some mental function like attention and memory (see Table 1), or SSRIs (see Table 2). For cognitive enhancers we took into account some of the drugs that are anecdotally reported as being popular among the student population [20]. In the case of the SSRIs we analysed widely prescribed drugs identified by previous studies [21]. In both cases we read the existing articles available at Wikipedia on each of the target drugs and obtained a list of synonyms for these drug names as shown in Tables 1 and 2.

2.1. First stage annotation

We used the Twitter streaming API [22] to obtain a random sample from all public tweets for a 12 month period (8th May 2012–20th April 2013). This gave us 420,983,674 messages. These data allowed us to understand how Twitter users mention the drugs of interest against a standard background.

Once the full random sample was gathered we used our synonym list to identify tweets mentioning any of the drugs of interest (see Tables 1 and 2). We then applied a further filter where we would only keep a maximum of 300 matching tweets (selected at random among the matched tweets) for each one of the 11 drugs, aiming at a maximum of 3300 tweets. This was done after we noticed that some drugs such as Adderall and Prozac had a far higher number of mentions than the other drugs. In order to obtain a balanced sample we set that upper bound of 300 samples for each drug. Moreover, in the case of *"Adrafinil"* we did not get a single mention on any of the synonyms we used. This can be considered an important finding on the sensitivity of the data source. The final data set used for our study consisted of 1548 tweets (see Tables 1 and 2). Since the distribution of drug mentions is not evenly balanced we will investigate a targeted approach in the future in order to increase the volume of rare drug name mentions. With the data in hand we constructed our gold standard annotation set by selecting 496 tweets to be annotated by 2 PhD students with training in computational linguistics (including the first author).

In order to check for influences on reporting bias we looked for popular stories that appeared during the time frame when we collected the tweets to check possible environmental influences from the media. The stories we found were *"FDA warns of counterfeit Adderall"* [23], *"John Moffitt on Adderall: 'It was a total mistake' "* [24], and *"Aurobindo Pharma gets USFDA nod for Modafinil tablets"* [25]. But on the whole there was no major evidence showing that these would have an impact on the data set we collected during the sample period.

The annotation categories we used were:

- **Tweet written in English language?** This question reported which tweets were written in English language.
- **Tweet about the drugs of interest?** Some drug names appeared as strings within the tweet, providing texts that were not of interest to us.
- **First-hand experience**: Used to identify personal use of the drug.
- **Other's Experience**: Used to identify someone else's use of the drug.
- **Activism**: Used to identify an alarm or call for change in the drug policy.
- **Cultural reference**: Used to identify when the annotator found the tweet referring to a song lyric, movie title, etc.
- **Humor**: Used to indicate that a tweet contained a formulaic joke, bumper sticker, etc.