Journal of Biomedical Informatics 58 (2015) 280–287

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

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*

ARTICLE INFO

Article history:
Received 21 October 2014
Revised 4 November 2015
Accepted 5 November 2015
Available online 7 November 2015

Keywords:
Crowdsourcing
Pharmacovigilance
Twitter
Natural language processing

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.

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

[*] Corresponding author at: Department of Informatics, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan.

E-mail addresses: nestoralvaro@nii.ac.jp (N. Alvaro), mike.conway@utah.edu (M. Conway), Son.Doan@kp.org (S. Doan), lofi@ifis.cs.tu-bs.de (C. Lofi), john.overington@stratifiedmedical.com (J. Overington), nhc30@cam.ac.uk (N. Collier).

http://dx.doi.org/10.1016/j.jbi.2015.11.004
1532-0464/© 2015 Published by Elsevier Inc.

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.

Table 1
Cognitive enhancers used in our study by drug name along with each synonym and number of tweets.

Drug name	Synonyms	# tweets
Adderall	Amphetamine mixed salts; amphetamine salt	300
Ritalin	Concerta; Daytrana; Phenida; Attenta; Hynidate; Focalin;	300
Modafinil	Modafinilum; Provigil; Sparlon; Alertec; Modavigil;	59
Adrafinil	Olmifon;	0
Armodafinil	Nuvigil	3

Table 2
SSRIs used in our study by drug name along with each synonym and number of tweets.

Drug name	Synonyms	# tweets
Citalopram	Celexa	65
Escitalopram	Lexapro; Cipralex	145
Paroxetine	Paxil; Seroxat	123
Fluoxetine	Prozac	300
Fluvoxamine	Luvox	14
Sertraline	Zoloft; Lustral	239

- **News**: Used to identify news items.
- **Info/resource**: Used to identify factoids or informational resources.
- **Marketing**: Used to identify sales of the drug product/accessory.
- **Opinion**: Used when the writer was reporting a personal opinion related to the drug.
- **Sentiment**: Used to describe whether the author was positive, negative or neutral in terms of sentiment about the drug.
- **Pleasure**: Used to indicate that the writer reports the drug usage as a pleasurable activity.
- **Craving**: Used to indicate that the writer reports stress relief related to the usage of the drug.
- **Disgust**: Used to indicate that the writer sees the studied drug usage or the drug users as repulsive.

For the initial annotation effort, we obtained the Cohen's Kappa [26] and Fleiss' Kappa [27] values comparing the inter-annotator agreement between experienced annotators as shown in Table 4 (columns 2 and 3) by using R's irr package [28]. We studied the Kappa values and identified possible causes of disagreement. These were loosely classified as follows:

- **Lack of context:** Some tweets were written using only proper and common nouns making it hard for the annotator to understand the tweet and whether the tweet was written in English. For example, *"@<PERSON> Ronaldo"*, *"@<PERSON> @<PERSON> @<PERSON> <DRUG> #rx" or "@<PERSON> <DRUG> FTW."* Major causes of disagreement were identified specifically in short tweets, the use of acronyms, emoticons, popular names and multilingual keywords.
- **Meaningless mention:** As the tweets were extracted based on keywords that matched the drug of interest's name it was very important to read the tweet carefully to confirm that the drug itself was mentioned, especially given that some user names in Twitter can resemble the drug name, e.g. *"@Adderall_RB I'm on it"*, *"RT @Adderall_XR: SO excited for the #entouragemovie"*. Here we can see how drug names do not appear in the tweets once we remove the user names (*"@<PERSON> I'm on it!"* and *"RT @<PERSON>: SO excited for the #entouragemovie"*, respectively).
- **Identifying first-hand reports:** We found that in some cases it was not straightforward to distinguish a first-hand experience from rhetorical thought: *"I wish I could prescribe <DRUG> myself for all these depressing ass tweets cheer tf up"*, and also how to annotate the tweet in the case of forwarding a tweet from someone else (doing a Retweet): *"RT @<PERSON>: @<PERSON> @<PERSON> – Fear not! I've got a couple of bottles of #<DRUG> right here. Pass me a doughnut, plea ..."*. In other cases it was not easy to tell for sure whether the writer was actually taking the drug: *"Popular antidepressants <DRUG>, <DRUG> and <DRUG> can lower libido and prevent orgasms #fact"*. In the same way it is not straightforward to realize whether the user took the drug and stopped taking it or whether she still takes it as in the following example: *"@<PERSON> yep. i honestly think the <DRUG> has messed up my memory and concentration or something because they suck now"*, *"Hello, <DRUG>. Miss me?"*.
- **Ambiguous genre:** Another area of disagreement was when annotating *"Opinions"* and *"Other's experience"*, as in some examples it could be understood in either way as in: *"@<PERSON> go to sleep already Joe and put down the <DRUG> really shit!"*, *"@<PERSON> @<PERSON> I just found it funny that people used <DRUG> against him."*, *"Jesse needs to lay off the <DRUG> lmao"*.

Our annotation guidelines for laymen and experienced annotators (included in *"Supplements"* file) elaborate on the basic questions shown in Table 4.

2.2. Crowdsourcing annotation

Although the two PhD annotators could have annotated all the tweets within the data set, given that experienced annotators are a scarce resource we decided to study other possibilities and rely on a crowdsourcing engine, also taking into account that the annotations obtained from the experienced annotators could be used as the gold standard when collecting laymen annotations.

We opted for CrowdFlower as the service allowed us to use a subset of the tweets previously tagged by our experienced annotators, enabling us to provide a set of data items with correct responses, which in turn were used to discard tainted contributions. We also configured the settings to target contributors from several English speaking countries (Australia, Canada, New Zealand, the United Kingdom, and the United States) on the assumption that annotators from these countries were more likely to be native English speakers.

We decided that the gold standard to be used in the crowdsourcing platform would be composed of 100 tweets where both expert annotators agreed on all fields. After that selection, the annotations provided by the expert annotators were then analysed by N.A. and N.C. to understand the cause of disagreements observing the points presented in the previous section. These 100 gold questions became the testing questions for laymen in CrowdFlower, acting as a filter to discard all the annotations coming from any annotator scoring lower than 70% on those test questions.

The experienced annotators used the extended version of the guidelines prior to annotation (Supplement 1: Expert annotator guidelines). These guidelines were based on those created for a study into usage of electronic tobacco products reported on social media [29]. All the categories in our study except three were also used in the electronic tobacco product study. We added two categories in order to refine the results by annotating whether the tweet was written in English, and also to focus on the drug reporting tweets. The third category we added was used to understand if the tweet was reporting a first hand experience.

Laymen annotators were presented with a simplified set of the annotation guidelines (Supplement 2: Laymen annotator guidelines) in the form of a questionnaire.

Once we obtained the aggregated results from CrowdFlower[1] we extracted the tweets that were written in English language and mentioned drugs of interest. This yielded 899 tweets that became our gold standard.[2]

3. Methods

Once we had the set of gold standard tweets we divided them randomly into training (2/3) and testing (1/3) sets, with 600 and 299 tweets respectively. The whole process is depicted in Fig. 1.

We found 356 tweets classified as first-hand experience tweets in the gold standard. This is 39.6% of the 899 tweets.

Having the data in place, we generated the features: n-grams, latent topics, orthographic features and other Twitter specific features (see Fig. 2 example). We used several linguistic feature types including character 1,2,3 -grams, e.g. 'za','oz'; word tokens, e.g. 'dies', bucketed message length in tokens, e.g. 10–20; topics (topic1, topic2...); and Twitter specific features (to check whether the tweet is addressed to someone by using the "@" sign, to check whether the tweet may want to stress something in particular by using the "#" sign...).

Previous research [30] showed that combining n-grams with other semantic features improves classification accuracy. In our approach, we did not use the raw n-grams (character n-grams, unigrams and bigrams). Instead we applied term frequency inverse document frequency (TF-IDF) weighting first.

There is clear evidence that LDA topic models provide valuable data with large text corpora and we decided to add it to our study based on recent studies that have shown its value for collections of Twitter messages [31,32]. The topics were discovered using Latent Dirichlet Allocation [33] on the training set, and as we had 11 groups of drugs but one of them had no matches (Table 1) we selected 40 topics corresponding to an even distribution across the tweets. We experimented with different number of topics (from 35 to 45), but the information gain method consistently reported that the LDA topics did not contribute as features. Our intention here was to investigate whether there is evidence that automated topic modelling could improve the accuracy of our system. If this is confirmed a natural next step would be to provide a semantic label for each topic [34].

After generating all the features we applied information gain as the feature reduction algorithm to obtain the best ranking features, given that it has superior performance over other feature reduction methods [35]. At that point we observed that the topics were not contributing as well as expected from previous studies [36], and we decided to concentrate on the top-ranking features discarding the use of LDA topics. We also observed that the bigrams were not listed as top-ranking features. This can be explained because our word bigrams had low frequency counts and indicates that it is better to focus on character n-grams where frequency is higher.

We used R's FSelector package [37] to calculate information gain. The algorithms finds weights for discrete attributes based on their correlation with the continuous class attribute. The formula it uses is the following, where it takes into consideration the entropies (represented by "*H*") of the class and the attribute:

$$\begin{aligned} \mathit{Information\ Gain} = {} & H(\mathit{Class}) + H(\mathit{Attribute}) \\ & - H(\mathit{Class}, \mathit{Attribute}) \end{aligned} \tag{1}$$

As shown in Table 3 we applied information gain to obtain the most discriminating 1%, 3%, 5%, 7%, 10%, 50%, and 100% features. When using 10% features we fed our models with 637 features. A sample of the obtained features is presented in Table 3.

4. Results

We used the "Full report" CrowdFlower provided, which contains all the annotations obtained from the contributors, to calculate the inter-annotator agreement showed in Table 4 (column 4, "*Fleiss' kappa for 5 raters (CrowdFlower)*"). The results were of comparable quality to the experienced annotators, although in general the crowdsourced results scored slightly lower than those obtained from experienced annotators. In the case of "*Activism*", "*News*", "*Marketing*" and "*Disgust*" the Kappa scores were higher than the values obtained from PhD raters. Once we obtained Fleiss' kappa results we ranked these values to calculate Spearman's Rho [38] (Rho = 0.471) and Kendall's tau [39] (0.352), where we observed moderate agreement [40]. This confirmed that the data we obtained from the crowdsourced annotations were of comparable quality to those obtained by expert annotators, a result consistent with previous work in the domain [41]. We observed several categories of question such as "*Cultural reference*" where the correlation values were markedly low. This is not surprising since Twitter contains many culture-specific references.

We further analysed the annotations from expert annotators to obtain Wilson score interval as suggested by [42]. Apart from calculating Wilson score interval between expert annotators we also computed the percentage agreement. The results are presented in Table 5.

CrowdFlower provided us with the "*aggregated*" results file, which only contains the most trustworthy annotation based on individual contributors' trust ratings for every question independent of the number of judgements that were requested per question (we requested 5 judgements per tweet). The confidence score describes the level of agreement between multiple contributors (weighted by the contributors' trust scores), and indicates CrowdFlower's "*confidence*" in the validity of the result [43]. Once a job is complete, all of the judgements on a row of data are aggregated with a confidence score, and in order to provide the aggregated result CrowdFlower chooses the response with the greatest confidence [44]. We used those "*aggregated*" results to train the machine learning models.

In order to control quality we had to apply some validation mechanism. We used expert annotators as stated in the Data Selection section to gauge laymen annotators quality. Apart from the validation mechanism we also believe it is important to mention the following points:

- **Resource scarcity**: Finding expert annotators was much harder than we initially expected. This, in the end, delayed the start of the experiments.
- **Costs**: Expert annotators were much more expensive to hire than laymen annotators. Given the experimental set up this point in particular did not affect us, but we realized it could have been an issue to consider in case we would have had to annotate a large amount of tweets.
- **Time constraints**: Expert annotators can only devote a limited number of hours per day to the annotation task. On the other hand, crowdsource annotators are a potentially unlimited work force and once the task was launched in CrowdFlower platform laymen annotators worked on it at a constant rate.

[1] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/1548_CrowdFlower.

[2] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/899_CrowdFlower.

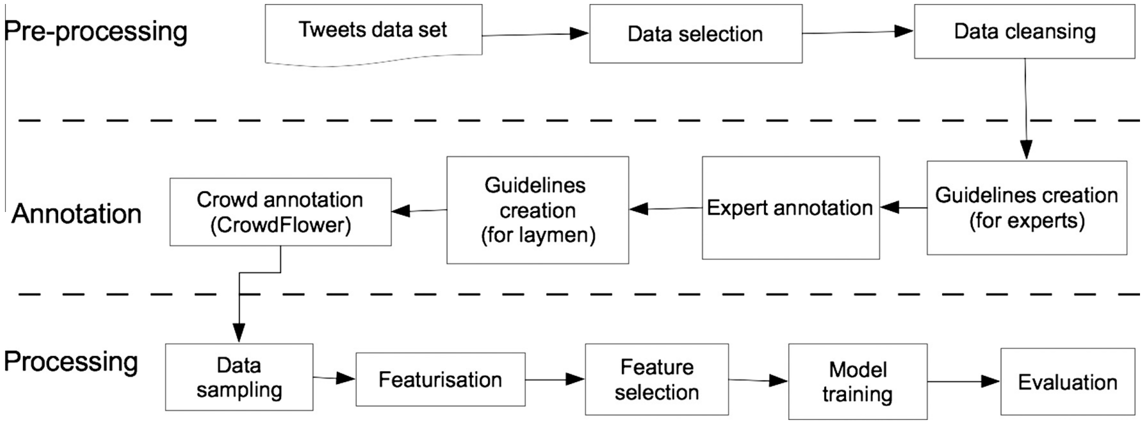

Fig. 1. Flowchart detailing the phases in our study. In the Pre-processing phase we obtain data from Twitter, extracted the tweets mentioning the drugs, and performed data cleansing. In the Annotation phase we performed both annotations (by experienced annotators and by laymen annotators). In the Processing phase we perform the featurisation and feature selection, used to train the models. Finally, the obtained models are evaluated.

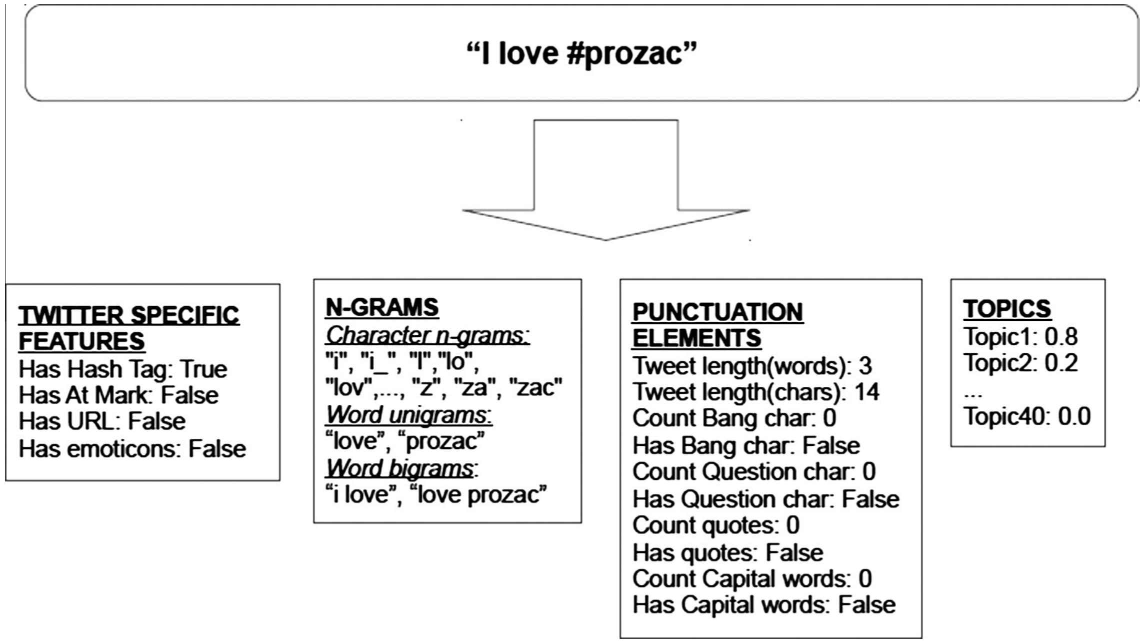

Fig. 2. Example of a featurised tweet including n-grams, latent topics, orthography and hashtags. Here we show the values of some of the features. In the case of the N-Grams we only show the obtained n-grams for such tweet.

Table 3
Sample of extracted features using 10% information gain.

Ranking	List of features
Top 10 features (features ranked 1–10)	*"my"*, *"za"*, *"zac"*, *"oza"*, *"roz"*, *"oz"*, *"ric"* (Character n-grams); *"Has hash tag"*, *"Has at mark"*, *"Has emoticons"*
10 features at the middle of the list (features ranked 313–322)	*"fill"*, *"filming"*, *"find"*, *"finding"*, *"findworkfamilylifebalance"*, *"fine"*, *"firsttestofthesemester"*, *"flip"*, *"flowing"*, *"flvs"* (unigrams)
10 features at the bottom of the list (features ranked 628–637)	*"phenergan"*, *"phenidaad"*, *"phillywcwagon"*, *"phoebebuffay"*, *"pib"*, *"pill"*, *"pizza"*, *"placenta"*, *"planet"*, *"plenty"* (unigrams)

After these observations we consider that both laymen and expert annotators contributed to our annotations very positively. We believe that the combination of laymen annotators, who can work on large volumes of data, and expert annotators, who can validate the annotations produced by laymen, provided a very good data set suited to our needs.

We then trained and tested C50, SVM using a linear kernel (SVM), Naive Bayes (NB), Multi-Layer Perceptron (MLP), Generalized Linear Model (GLM), and Bayesian Generalized Linear Model (BGLM) from R's Caret package [45] to assess their performance on our data sets using the selected set of features.

Table 4
Inter annotator agreement between raters using Cohen's and Fleiss' Kappas.

Question	Cohen's kappa for experienced raters	Fleiss' kappa for experienced raters	Fleiss' kappa for 5 raters (CrowdFlower)
Tweet written in English language?	0.962	0.962	0.943
Tweet about the drugs of interest?	0.888	0.888	0.845
First-hand experience	0.674	0.673	0.556
Other's Experience	0.391	0.390	0.231
Activism	−0.002	−0.005	0.075
Cultural reference	0.427	0.424	0.112
Humor	0.392	0.390	0.377
News	0.338	0.336	0.352
Info/resource	0.382	0.381	0.294
Marketing	0.361	0.357	0.409
Opinion	0.282	0.266	0.244
Sentiment	0.395	0.385	0.314
Pleasure	0.076	0.075	0.057
Craving	0.362	0.360	0.239
Disgust	0.045	0.044	0.129

Table 5
Wilson confidence interval (minimum and maximum), and percentage agreement between 2 expert annotators.

Question	Wilson conf. interval (min)	Wilson conf. interval (max)	Percentage agreement
Tweet written in English language?	0.968	0.990	0.982
Tweet about the drugs of interest?	0.925	0.960	0.945
First-hand experience	0.876	0.922	0.902
Other's experience	0.920	0.956	0.941
Activism	0.978	0.995	0.989
Cultural reference	0.945	0.974	0.962
Humor	0.876	0.922	0.902
News	0.963	0.986	0.977
Info/resource	0.907	0.946	0.929
Marketing	0.959	0.984	0.974
Opinion	0.847	0.897	0.874
Sentiment	0.850	0.900	0.877
Pleasure	0.954	0.980	0.970
Craving	0.948	0.977	0.965
Disgust	0.963	0.986	0.977

For the evaluation of the results we use the *F*-Score, based on the standard precision and recall:

$$F\text{-}Score = 2 * \frac{precision * recall}{precision + recall} \tag{2}$$

where recall is:

$$Recall = \frac{true\ positives}{true\ positives\ +\ false\ negatives} \tag{3}$$

And precision is:

$$Precision = \frac{true\ positives}{true\ positives\ +\ false\ positives} \tag{4}$$

To better quantify the performance of the models we also include the Informedness measures [46]. The Informedness measure, apart from taking into account the *"true positive"*, *"false positive"* and *"false negative"* values that are used by the *F*-Score, uses the *"true negative"* values getting a fair measure for classification showing which are the most informative models and which are the models that even when obtaining high *F*-Score values do not have predictive power.

$$Informedness = recall + invRecall - 1 \tag{5}$$

where inverse recall is:

$$invRecall = \frac{true\ negatives}{true\ negatives\ +\ false\ positives} \tag{6}$$

4.1. First evaluation using the initial data set

As shown in Table 6, combining the six learning models with the selected set of features gave a maximum *F*-Score of 0.64 when using CrowdFlower data. BGLM is the best performing model, followed by C50. GLM is the other model scoring above the baseline. The baseline, obtained by predicting all labels to be *"First-hand experience"*, achieves an *F*-score of 0.55. In this and following experiments the *"NaN"* value in the tables indicate that all predicted labels were *"Other genre"*. Here we can see that BGLM is the most informative model, followed by GLM.

4.2. Second evaluation using the initial data set

For our next experiment we asked an expert annotator (N.A.) to annotate the fields *"First-class experience"*, *"Tweet written in English language"*, and *"Tweet about the drug"* for the same 1548 tweets that the laymen annotated. After having these annotations we discarded all the tweets where the laymen and the expert disagreed on the annotation for those fields, obtaining the 661 tweets[3] that we used to run the same experiment from before. We present the results from this experiment in Table 7. In this case the baseline is also obtained when labelling all tweets as *"First-hand experiences"* and has 0.45 *F*-Score. In this experiment BGLM was the best model both in terms of *F*-score and Informedness.

4.3. Extended evaluation

During September 26th 2014 until December 9th 2014 we collected a new data set from Twitter by filtering the tweets containing any of the drug names or drug synonyms listed in Tables 1 and 2. We gathered 159,007 tweets and chose 4000 tweets at random to be annotated by two experts using the same version of the guidelines. We obtained 3211 tweets where both expert annotators agreed on the annotation for the genre and which were written in English language and about the drugs of interest. We used that dataset[4] as the gold standard for our last experiment.

In this experiment we obtained a much larger number of feature values, and in order to process all of them (mainly because of computer memory limitations) we had to reduce the number

[3] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/661_CrowdFlower_Expert.

[4] A modified version of this file complying with Twitter's TOS can be found on github https://github.com/nestoralvaro/JBI_Pharmacovigilance/tree/master/3211_Experts.

Table 6
F-score values for each model using a selected percentage of features on 899 tweets annotated via crowdsourcing. Note that figures in parentheses show the Informedness values. The highest values in each column are highlighted in bold.

Model	1%	3%	5%	7%	10%	50%	100%
SVM	0.48 *(0.15)*	0.49 *(0.20)*	0.48 *(0.18)*	0.46 *(0.15)*	0.43 *(0.13)*	0.40 *(0.12)*	0.28 *(0.00)*
C50	0.61 *(0.27)*	0.61 *(0.27)*	**0.61** *(0.27)*	0.61 *(0.27)*	0.61 *(0.27)*	**0.57** *(0.10)*	**0.55** *(0.00)*
GLM	0.59 *(0.38)*	0.57 *(0.35)*	0.54 *(0.32)*	0.53 *(0.33)*	0.56 *(0.32)*	0.40 *(−0.06)*	0.42 *(0.01)*
MLP	0.47 *(0.11)*	0.52 *(0.24)*	0.42 *(0.11)*	0.37 *(0.04)*	0.44 *(0.03)*	0.44 *(0.07)*	0.47 *(0.11)*
BGLM	**0.64** **(0.43)**	**0.63** ***(0.41)***	**0.61** ***(0.39)***	**0.64** ***(0.43)***	**0.62** ***(0.40)***	0.55 ***(0.28)***	0.54 ***(0.27)***
NB	0.13 *(0.04)*	0.10 *(0.03)*	0.02 *(0.00)*	NaN *(0.00)*	NaN *(0.00)*	NaN *(0.00)*	NaN *(0.00)*

Table 7
F-score values for each model using a selected percentage of features on 661 tweets annotated via crowdsourcing and by an expert. Note that figures in parentheses show the Informedness values. The highest values in each column are highlighted in bold.

Model	1%	3%	5%	7%	10%	50%	100%
SVM	0.15 *(−0.03)*	0.09 *(−0.07)*	0.14 *(0.01)*	0.13 *(−0.03)*	0.09 *(−0.07)*	0.27 *(0.08)*	0.20 *(0.04)*
C50	0.24 *(0.13)*	0.52 *(0.32)*	0.39 *(0.22)*	0.28 *(0.15)*	0.43 *(0.26)*	0.48 *(0.14)*	0.47 *(0.17)*
GLM	0.50 *(0.32)*	0.37 *(0.19)*	0.30 *(0.15)*	0.30 *(0.15)*	0.35 *(0.17)*	0.30 *(−0.15)*	0.36 *(0.01)*
MLP	0.19 *(−0.06)*	0.25 *(−0.03)*	0.25 *(0.03)*	0.21 *(−0.01)*	0.32 *(0.08)*	0.39 *(0.17)*	0.27 *(0.01)*
BGLM	**0.57** ***(0.40)***	**0.55** ***(0.37)***	**0.56** ***(0.39)***	**0.51** ***(0.33)***	**0.50** ***(0.31)***	**0.57** ***(0.40)***	**0.57** ***(0.39)***
NB	0.21 *(0.09)*	NaN *(0.00)*	NaN *(0.00)*	NaN *(0.00)*	NaN *(0.00)*	NaN *(0.00)*	0.41 *(−0.03)*

Table 8
F-score values for each model using a selected percentage of features on 3211 tweets annotated by two experts. Note that figures in parentheses show the Informedness values. The highest values in each column are highlighted in bold.

Model	1%	3%	5%	7%	10%	50%	100%
SVM	0.58 *(0.26)*	0.49 *(0.15)*	0.29 *(0.04)*	0.35 *(0.05)*	0.48 *(0.14)*	0.55 *(0.14)*	0.27 *(0.04)*
C50	0.21 *(0.09)*	0.14 *(0.06)*	0.11 *(0.05)*	0.15 *(0.08)*	0.29 *(0.16)*	**0.75** ***(0.57)***	0.09 *(0.04)*
GLM	**0.77** ***(0.59)***	0.75 ***(0.57)***	0.75 *(0.56)*	0.74 *(0.54)*	0.68 *(0.47)*	0.36 *(0.02)*	0.48 *(0.01)*
MLP	0.63 *(0.33)*	0.65 *(0.35)*	0.54 *(0.11)*	0.53 *(0.14)*	0.56 *(0.19)*	NaN *(0.00)*	0.56 *(0.20)*
BGLM	**0.77** ***(0.59)***	**0.76** ***(0.57)***	**0.76** ***(0.57)***	**0.77** ***(0.59)***	**0.77** ***(0.59)***	0.68 *(0.41)*	**0.77** ***(0.59)***
NB	0.63 *(0.09)*	0.62 *(0.06)*	0.64 *(0.14)*	0.66 *(0.24)*	NaN *(0.00)*	NaN *(0.00)*	NaN *(0.00)*

of character n-grams and only keep those that appeared more than ten times. Apart from this change, the code we used for training and testing was the same that we used when we ran the experiments reported in the previous sections.

We present in Table 8 the *F*-Score results obtained for each model and each set of features. In this dataset the baseline prediction is obtained when labelling all tweets as *"First-hand experience"* tweets (0.61 *F*-Score). Here BGLM gets the highest *F*-Score and Informedness results for almost all sets of features.

5. Conclusions

In this paper we explored the classification of Twitter messages into first-hand drug user experience. For the task of selecting ADR data on the crowdsourced annotations Bayesian Generalized Linear Model (BGLM) was observed to be the model providing the overall highest *F*-Score among those tested, only surpassed by C50 when using the top 50% and the 100% of the features, although in terms of Informedness BGLM obtained the best scores all the time.

We also used the subset of the same data for which both the laymen and one expert agreed on the annotation for the fields *"First-class experience"*, *"Tweet written in English language"*, and *"Tweet about the drug"*. In this case BGLM obtained the best *F*-Score values, and also the highest Informedness measure, showing the predictive power of this model for this dataset.

For our last experiment we used the dataset where the annotations from two expert annotators were in agreement for the fields *"First-class experience"*, *"Tweet written in English language"*, and *"Tweet about the drug"*. In this experiment we observed that BGLM had the highest *F*-Score values, only matched by GLM when using the top 1% features. This is particularly interesting because the annotators were not laymen, and the data were collected during a different period and also using a different method, but the best performing model was the same as in the previous experiments.

We also observed that most models had a stable performance independent of the set of features. We also realized that *"SVM"* predictions were lower than the baseline in all the experiments, and *"Multi-Layer Perceptron"*, and *"Naive Bayes"* only scored above the baseline when using the dataset annotated by the two experts.

We believe this line of research can be meaningful given the volume of tweets that are constantly generated. Having a first filter to detect user reports on Twitter on the drug use can help in pruning valuable data since the beginning of other studies. Our aim is to continue exploring this path to automatically identify tweets reporting first-hand experiences on a set of drugs, and our plan is to further study how a feature re-engineering process should be performed, in particular when combined with LDA topics and ensemble models. We will also consider expanding the list of synonyms to include slang names for the selected drugs.

Importantly, we also showed that the inter-annotator agreement from CrowdFlower is of comparable quality to the inter-annotator agreement obtained from experienced annotators, confirming that we can rely on crowdsourced annotations to identify personal drug reports, although there are still difficulties such as some notable disagreements (e.g. cultural references, disgust) that need to be recognised. To overcome this we have to analyse how human agreement might be improved as there are some open areas of work such as better guideline development and better interface selection.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research project was supported by a grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). Nigel Collier was supported by a Marie Curie fellowship (PI: Phenominer, Grant No. 301806). We would also like to thank Dr. Danielle Mowery, University of Utah for helping in annotating the tweets. Finally, we want to thank Pascual Martínez-Gómez for useful discussions on the software implementation for evaluation, and also thank the anonymous reviewers who helped us in improving the quality of the paper with their comments and suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.11.004.

References

[1] L.T. Kohn, J.M. Corrigan, M.S. Donaldson, et al., To Err is Human: Building a Safer Health System, vol. 627, National Academies Press, 2000.

[2] J. Lazarou, B.H. Pomeranz, P.N. Corey, Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies, JAMA: J. Am. Med. Assoc. 279 (15) (1998) 1200–1205.

[3] The International Conference on Harmonisation Expert Working Group, Guideline for Good Clinical Practice e6(r1), 1996. <http://private.ich.org/LOB/media/MEDIA482.pdf>.

[4] Office of the Commissioner, Medwatch: The FDA Safety Information and Adverse Event Reporting Program, 2014. <http://www.fda.gov/Safety/MedWatch/default.htm>.

[5] L. Hazell, S.A. Shakir, Under-reporting of adverse drug reactions, Drug Saf. 29 (5) (2006) 385–396.

[6] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks, in: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, 2010, pp. 117–125.

[7] C.L. Hanson, B. Cannon, S. Burton, C. Giraud-Carrier, An exploration of social circles and prescription drug abuse through Twitter, J. Med. Internet Res. 15 (9) (2013).

[8] FDA, Guidance for Industry Internet/Social Media Platforms with Character Space Limitations Presenting Risk and Benefit Information for Prescription Drugs and Medical Devices, June 2014. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm401087.pdf>.

[9] Medicines and Healthcare Products Regulatory Agency (MHRA). Press Release: UK Regulator Leads Innovative EU Project on the use of Smartphones and Social Media for Drug Safety Information, November 2015. <http://www.imi.europa.eu/content/web-radr>.

[10] E.M. Agency, European Medicines Agency. Guideline on Good Pharmacovigilance Practices (GVP), 2013. <http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000345.jsp>.

[11] E.M. Agency, Guideline on Good Pharmacovigilance Practices (GVP) Module VI Management and Reporting of Adverse Reactions to Medicinal Products (rev 1), 2014. <http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf>.

[12] United States Securities and Exchange Commission, Form s-1. Registration statement. Twitter, inc., October 2013. <http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm>.

[13] A. Nikfarjam, G.H. Gonzalez, Pattern mining for extraction of mentions of adverse drug reactions from user comments, AMIA Annual Symposium Proceedings, vol. 2011, American Medical Informatics Association, 2011, p. 1019.

[14] F. González-Rubio, A. Calderón-Larrañaga, B. Poblador-Plou, C. Navarro-Pemán, A. López-Cabañas, A. Prados-Torres, Underreporting of recognized adverse drug reactions by primary care physicians: an exploratory study, Pharmacoepidem. Drug Saf. 20 (12) (2011) 1287–1294.

[15] Revealed: The Demographic Trends for Every Social Network, October 2014. <http://www.businessinsider.com/2014-social-media-demographics-update-2014-9>.

[16] S. Schneeweiss, A.R. Patrick, D.H. Solomon, C.R. Dormuth, M. Miller, J. Mehta, J. C. Lee, P.S. Wang, Comparative safety of antidepressant agents for children and adolescents regarding suicidal acts, Pediatrics (2010). peds–2009.

[17] C.L. Hanson, S.H. Burton, C. Giraud-Carrier, J.H. West, M.D. Barnes, B. Hansen, Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students, J. Med. Internet Res. 15 (4) (2013).

[18] C.C. Freifeld, J.S. Brownstein, C.M. Menone, W. Bao, R. Filice, T. Kass-Hout, N. Dasgupta, Digital drug safety surveillance: monitoring pharmaceutical products in Twitter, Drug Saf. 37 (5) (2014) 343–350.

[19] U. Food, D. Administration, FDA Adverse Event Reporting System (FAERS), 2015. <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/>.

[20] C.I. Ragan, I. Bard, I. Singh, What should we do about student use of cognitive enhancers? An analysis of current evidence, Neuropharmacology 64 (2013) 588–595.

[21] N.C. White, T. Litovitz, C. Clancy, Suicidal antidepressant overdoses: a comparative analysis by antidepressant type, J. Med. Toxicol. 4 (4) (2008) 238–250.

[22] Twitter Inc., Get Statuses/Sample, September 2014. <https://dev.twitter.com/streaming/reference/get/statuses/sample>.

[23] D. Sell, FDA Warns of Counterfeit Adderall, October 2012. <http://articles.philly.com/2012-05-31/business/31900817_1_rogue-websites-and-distributors-generic-versions-adderall>.

[24] D. O'Neil, John Moffitt on Adderall: 'it was a total mistake', November 2012. <http://seattletimes.com/html/seahawksblog/2019783660_adderall28.html>.

[25] Aurobindo Pharma Gets USFDA nod for Modafinil Tablets, September 2012. <http://articles.economictimes.indiatimes.com/2012-09-28/news/34148290_1_aurolife-pharma-llc-usfda-nod-aurobindo-pharma>.

[26] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Measur. 20 (1) (1960) 37–46.

[27] J.L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educ. Psychol. Meas. (1973).

[28] M. Gamer, Package 'irr', 2012. <http://cran.r-project.org/web/packages/irr/irr.pdf>.

[29] M. Myslín, S.-H. Zhu, W. Chapman, M. Conway, Using Twitter to examine smoking behavior and perceptions of emerging tobacco products, J. Med. Internet Res. 15 (8) (2013).

[30] M. Conway, S. Doan, A. Kawazoe, N. Collier, Classifying disease outbreak reports using n-grams and semantic features, Int. J. Med. Inform. 78 (12) (2009) e47–e58.

[31] Y. Tewari, R. Kawad, Real-Time Topic Modeling of Microblogs, March 2013. <http://www.oracle.com/technetwork/articles/java/micro-1925135.html>.

[32] R. Mehrotra, S. Sanner, W. Buntine, L. Xie, Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling, 2013.

[33] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[34] J.H. Lau, K. Grieser, D. Newman, T. Baldwin, Automatic labelling of topic models, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, 2011, pp. 1536–1545.

[35] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: ICML, vol. 97, 1997, pp. 412–420.

[36] L. Hong, B.D. Davison, Empirical study of topic modeling in Twitter, in: Proceedings of the First Workshop on Social Media Analytics, ACM, 2010, pp. 80–88.

[37] P. Romanski, Package "FSelector, February 2013. <http://cran.r-project.org/web/packages/FSelector/FSelector.pdf>.

[38] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1) (1904) 72–101.

[39] M.G. Kendall, A new measure of rank correlation, Biometrika (1938) 81–93.

[40] C. Dancy, J. Reidy, Statistics without maths for psychology, IEEE Statistics without maths for psychology, 2004.

[41] S. Nowak, S. Rüger, How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation, in: Proceedings of the International Conference on Multimedia Information Retrieval, ACM, 2010, pp. 557–566.

[42] L.D. Brown, T.T. Cai, A. DasGupta, Interval estimation for a binomial proportion, Stat. Sci. (2001) 101–117.

[43] How to Calculate a Confidence Score, July 2014. <http://success.crowdflower.com/customer/portal/articles/1295977-how-to-calculate-a-confidence-score>.

[44] CrowdFlower, Get Results – How to Calculate a Confidence Score, 2015. <https://success.crowdflower.com/hc/en-us/articles/201855939-Get-Results-How-to-Calculate-a-Confidence-Score>.

[45] M. Kuhn, Package 'caret', August 2014. <http://cran.r-project.org/web/packages/caret/caret.pdf>.

[46] D.M. Powers, Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation, 2011.