Analysis of Twitter Data for Postmarketing Surveillance in Pharmacovigilance

Julie Pain and **Jessie Levacher** and **Adam Quinquenel**
INSA Rouen
Avenue de l'Université, Saint-tienne-du-Rouvray, France
{julie.pain,jessie.levacher,adam.quinquenel}@insa-rouen.fr

Anja Belz
Computing, Engineering and Mathematics
University of Brighton, Lewes Road, Brighton, UK
a.s.belz@brighton.ac.uk

Abstract

Postmarketing surveillance (PMS) has the vital aim to monitor effects of drugs after release for use by the general population, but suffers from under-reporting and limited coverage. Automatic methods for detecting drug effect reports, especially for social media, could vastly increase the scope of PMS. Very few automatic PMS methods are currently available, in particular for the messy text types encountered on Twitter. In this paper we describe first results for developing PMS methods specifically for tweets. We describe the corpus of 125,669 tweets we have created and annotated to train and test the tools. We find that generic tools perform well for tweet-level language identification and tweet-level sentiment analysis (both 0.94 F1-Score). For detection of effect mentions we are able to achieve 0.87 F1-Score, while effect-level adverse-vs.-beneficial analysis proves harder with an F1-Score of 0.64. Among other things, our results indicate that MetaMap semantic types provide a very promising basis for identifying drug effect mentions in tweets.

1 Introduction

Postmarketing surveillance (PMS) is a vital part of pharmacovigilance, taking place during the final, post-licensing phase of drug development when the effects on larger numbers of users, who may have other conditions and take other medicines than those included in pre-approval tests, can be assessed. PMS is implemented in passive national reporting schemes such as the Yellow Card Scheme in the UK and MedWatch in the US; it is also implemented as active surveillance, e.g. organisations such as the MHRA in the UK and the FDA in the US conduct post-approval studies and postmarketing surveys.

Existing PMS methods either rely on health practitioners and patients to report adverse effects to which only a small (self-selected) proportion of patients in particular will contribute (reporting schemes), or they involve single products and small numbers of participants (surveys). More generally, such methods are more likely to identify (i) very serious problems, and (ii) problems relating to newly released drugs and drugs already under continuous surveillance.

The ultimate goal of our work is to develop text analysis techniques to facilitate automatic, continuous and large-scale monitoring of adverse drug reactions (ADRs), and more generally of effects, beneficial or otherwise, reported for given drugs. There is currently huge interest in such methods in the pharmaceutical industry, in particular where they can be used to monitor what is being said on social media about specific drugs. Automatic PMS methods are also of interest to national regulatory bodies.

The potential benefits of high-precision automatic ADR detection methods for social media are great: such methods would ameliorate the recognised problem of under-reporting of ADRs via existing channels (Lardon et al., 2015), and could inform the design of post-approval studies. Studies have already demonstrated that analysis of social media contents (manual analysis so far) can lead to the discovery of serious side effects, e.g. Abou

Proceedings of the 2nd Workshop on Noisy User-generated Text,
pages 94–101, Osaka, Japan, December 11 2016.

Taam et al. (2014) retrospectively identified a severe side effect on the basis of user content posted months before the drug in question was withdrawn because of the same side effect. The challenge is to automate this process so that suitably large portions of social media can be scanned.

2 Related Research

The main interest has so far been in discovering ADRs, defined by the World Health Organisation as a "a response to a medicinal product which is noxious and unintended and which occurs at doses normally used in [humans] for the prophylaxis, diagnosis or therapy of disease or for the restoration, correction or modification of physiological function" (WHO, No date).

While studies involving manual analysis have confirmed the usefulness of social media content in ADR identification (see previous section), manual analysis is necessarily limited to minute fractions of available online content. Automatic analysis of online content for ADR detection remains a huge challenge, in particular when applied to Twitter, due to the messy and 'un-language-like' nature of tweets, and it remains a small field.

Last year, Lardon et al. (2015) reported the results of a very thorough scoping review of (i) manual and computer-aided identification of ADRs, and (ii) semi-automatic and automatic identification of ADRs, from social media. The authors found just 13 papers on the latter, of which just one paper focused on Twitter data (the remainder used content from online health forums).

That one paper (Bian et al., 2012), among the very earliest on this precise topic, reported the creation and annotation of a corpus of drug-related tweets, and results for training SVM classifiers to detect (a) whether the tweeter was reporting their own experience, and (b) whether a given own-experience tweet contained a mention of an ADR. The study targeted five drugs that were the subject of a pre-approval clinical study during a known time window, and analysed tweets that were posted during the same time window, on the assumption that study participants were likely to tweet about their experience. 489 tweets by 424 users were identified in this way, reduced to 239 users after removal of re-tweets and non-English tweets. Tweets by the same user were concatenated and annotated manually. An SVM classifier was trained to classify texts into (i) users reporting their own experience of taking the drug, and (ii) user reporting someone else's experience; the classifier used 171 features, some based on keywords and word types, hashtags and user names; others based on a MetaMap (Aronson, 2001) analysis of the texts yielding UMLS meta codes. Classification into first-hand vs. second-hand reporting achieved an average accuracy of 0.74, and classification of first-hand-reporting tweets into ADR-mentioning and not ADR-mentioning was also reported as achieving 0.74 accuracy.

Ginn et al. (2014) created a corpus of annotated tweets and results for training classifiers to detect ADR mentions. Two members of the project team with "medical or biological science background" annotated 10,822 tweets related to 74 drugs for mentions of an ADR, indication or beneficial effect, and each ADR/indication/beneficial effect with corresponding UMLS concepts. The 74 drugs included some for which ADRs are well established and some for which they are not. Product names, common names and misspellings were used for each drug; tweets with URLs were removed (deemed to be mostly adverts). Inter-annotator agreement (IAA) was kappa = 0.69. Ginn et al. trained Naive Bayes (NB) and SVM classifiers on the binary annotations (presence/absence of effect mentions), using bag-of-words feature vectors, text normalisation and lemmatisation. Three versions of the corpus were created with different levels of imbalance between positive and negative examples. Best Precision for detecting effects was 0.89 (Recall = 0.695; F-Score=0.78; Accuracy = 0.746) for NB and the exactly balanced version of the corpus; Accuracy was around 0.75 for NB across the three differently balanced corpus versions.

Two members of the Ginn et al. team subsequently reported a larger set of results (Sarker and Gonzalez, 2015), using the Twitter data above, but also data from an online health forum, and an existing corpus of clinical reports. They describe the earlier results as having been obtained by "classification via under-sampling, which yields higher ADR F-scores at the expense of overall accuracy" (Footnote 22). Sarker & Gonzalez decided not to use under-sampling or balancing, and report ADR F-Scores of 0.538 and Accuracy 86.2 for the Twitter data with a different type of SVM than in the earlier work, going up to 0.597 and 90.1, respectively, when the Twitter data is supplemented by

the data from the online health forum.

In summary, the tasks for which tools have so far been built under the heading of automatic ADR detection in tweets are all binary (note that the *annotations* in the data sets include a wider variety of information):

- for a given tweet, decide whether it is in English or not (Bian et al., 2012);
- for a given set of tweets by the same Twitter user and a given drug, decide whether or not the user themself is taking (or has taken) the drug (Bian et al., 2012);
- for a given tweet, decide whether or not it contains a mention of an ADR (Bian et al., 2012; Ginn et al., 2014; Sarker and Gonzalez, 2015).

In this paper, we address (i) the first task above (Section 5), (ii) a more general case of the last task above (mentions of all drug reactions, adverse or not, Section 7), as well as two new, non-binary tasks (for this domain): (iii) for a given tweet, decide what its sentiment is (Section 6); and (iv) for a tweet that has been identified as containing a mention of a drug reaction (as in ii above), decide whether the drug reaction is beneficial, adverse or neither (Section 8).

3 Automatic Drug Effect Detection: Task, Annotation, Tools

Automatic drug effect detection for Twitter is something of a worst case scenario for NLP: not only are tweets one of the messiest, most abbreviated forms of short text, for which standard NLP tools do not tend to work well; but automatic ADR detection is also extremely hard, even for very well behaved types of text, due among other factors to the large variety of ways in which the same drug effects can be referred to, and to the complex relationships between drugs and effects.

The research reviewed in the previous section focused on detecting *adverse* drug effects (the annotations in some cases also include other kinds of effects, but the tools were for ADR detection). There are three reasons why we address a wider set of effects: (a) the related research area of automatic drug discovery/development, including our industrial advisory partner, is interested in identifying *all* effects claimed for a given drug/compound; (b) post-marketing drug monitoring, especially by pharmaceutical companies, is also interested in a wider set of effects; and (c) it may make the detection task harder if one specific subtype of drug effects is targeted only.

To elaborate on that last point, ADRs can be seen as a special case of a more general binary drug-effect relation *has_effect(drug, effect)*. Our corpus (Section 4) contains diverse examples, including the following:[1]

has_effect("glucosamine", "got me feeling some typa way")
has_effect("azathioprine", "think I've got #shingles")
has_effect("daunorubicin","made his wee red")

Reports of effects, whether positive, negative, or neither, are likely to have similar patterns and cue phrases (e.g. 'I've got', 'got me feeling'). It may be the case that it is easier to identify all reports of effects, not just negative ones, and then classify those into adverse, beneficial, mild, severe, and other dimensions.

We construe drug-effect detection as a knowledge extraction task, where the objective is to identify mentions of drug effects and to fill relation templates such as *has_effect(drug, effect)*. In this general form such mentions can be incorporated into knowledge graphs and combined with other kinds of relations about drugs and health. Such knowledge graphs can be easily visualised, and are used for example in automatic drug discovery research where mentions of drug effects from social media could provide a useful complementary source of information.

In this paper, we report first results towards the above knowledge extraction task. We use the term 'drug effect' to refer to both cases where the effect is specified and cases that might more intuitively be described as 'properties' where there is no specified effect (see also Section 8 below). Section 4 describes how we collected our set of tweets, and the text cleaning and normalisation we perform. Sections 5, 6, and 7 report results from our experiments on language identification, sentiment analysis and drug effect detection in tweets, respectively, while Section 8 reports results from classifying drug effects into beneficial, adverse and other.

[1] The text would be mapped to e.g. UMLS codes before being incorporated into knowledge bases, as below.

4 Data Collection

We compiled a list of drug names by extracting all names of approved drugs from DrugBank,[2] 1,999 in total. Next, for each drug name we collected the HTML files of all tweets returned for the search *"Drug Name" OR #DrugName* (e.g. *"Acetic Acid" OR #AceticAcid*), up to a maximum of 100 tweets per drug name, and used Beautiful Soup[3] to extract the information we needed (tweet text, tweet id, timestamp, etc.) from the HTML files containing the search results (tweets).

The results was a corpus[4] of 125,669 tweets, with an average of 62.87 tweets collected for each drug. Figure 1 gives a more detailed picture of number of tweets per drug: most drugs have 70-95 tweets, 19 have none, and 15 have 100 or more. Note that we did not expand the set of drug names with misspelt variants, generic or common names as done in some related research (Ginn et al., 2014).

We used three versions of the tweet texts: (i) in their raw form; (ii) cleaned, with URLs and user names replaced by tags, and hash tags with the '#' removed; and (iii) with hashtags additionally converted to likely strings of words, using a tool made available on StackOverflow.[5]

5 Language Identification

In order to be able to analyse tweets in a meaningful way, moreover using text analysis tools trained for English, it is important that we can reliably filter out non-English tweets. Twitter tags tweets for language, so part of our first set of experiments was to test the reliability of the Twitter language tags, as well as to see how it compared against existing language identification tools. For the latter we chose the three tools that were, combined as an ensemble method, identified by Lui & Baldwin (2014) as the best language identification tools for tweets: langid.py (Lui and Baldwin, 2012), LangDetect,[6] and CLD2.[7]

We randomly selected a subset of 300 tweets (test set A) and manually annotated each for English vs. other. Table 5 shows results for each of the methods tested, as well as for combining langid.py, CLD2 and LangDetect in majority voting. langid.py outperforms the others, including Twitter, slightly on the raw tweets (see end of Section 4). First cleaning/normalising tweets led to a slight improvement for CLD2, and a slight deterioration for languid.py; additionally parsing hashtags led to slight improvement for languid.py, and slight deteriorations for CLD2 and LangDetect. The final best result is an F1-Score of 0.94 for langid.py. This confirms (for this domain) previous results that good language identification tools outperform Twitter language tags (Lui and Baldwin, 2014).

	Twitter	langid.py	CLD2	LangDetect	Majority vote
P	0.995	0.991	0.986	0.973	0.995
R	0.861	0.889	0.877	0.898	0.861
F1	0.923	**0.937**	0.928	0.933	0.923

Table 1: Language identification results for Twitter tags and three language identification tools on test set A of 300 manually annotated tweets (Recall, Precision, F1-Score, for the English class).

6 Sentiment Analysis

Following the language identification experiments, we filtered out the non-English tweets using langid.py, which left us with 93,697 tweets and an average of 46.87 tweets per drug name. In the experiments in this section, we aim to predict the overall sentiment of an (English) tweet, independently of whether a drug effect is being reported.

For this purpose we selected a different test set of 300 random tweets (test set B) from our corpus and annotated each tweet for positive, negative, or neutral. We tested the 18 sentiment analysis (SA) tools made available in the ifeel package:[8] Afinn, Emolex, Emoticons, EmoticonDS, HappinessIndex, MPQAAdapter, NRCHashtagSentimentLexicon, OpinionLexicon, Panas-t, Sann, SASA, SenticNet, Sentiment140Lexicon, SentiStrength-Adapter, SentiWordNet, SoCal, UmigonAdapter, Vader, and a majority vote by all methods. The ifeel team have reported comparative results for many of these tools (Nunes Ribeiro et al., 2010; Ribeiro et al., 2016). For our task, the methods

[2] http://www.drugbank.ca/drugs

[3] https://www.crummy.com/software/BeautifulSoup

[4] We will publish the corpus along with this paper.

[5] By anonymous user Generic Human.

[6] Y. Nakatani: http://www.slideshare.net/shuyo/language-detection-library-for-java

[7] Language identification in Google Chrome.

[8] http://blackbird.dcc.ufmg.br:1210/ (accessed: 06/2016)

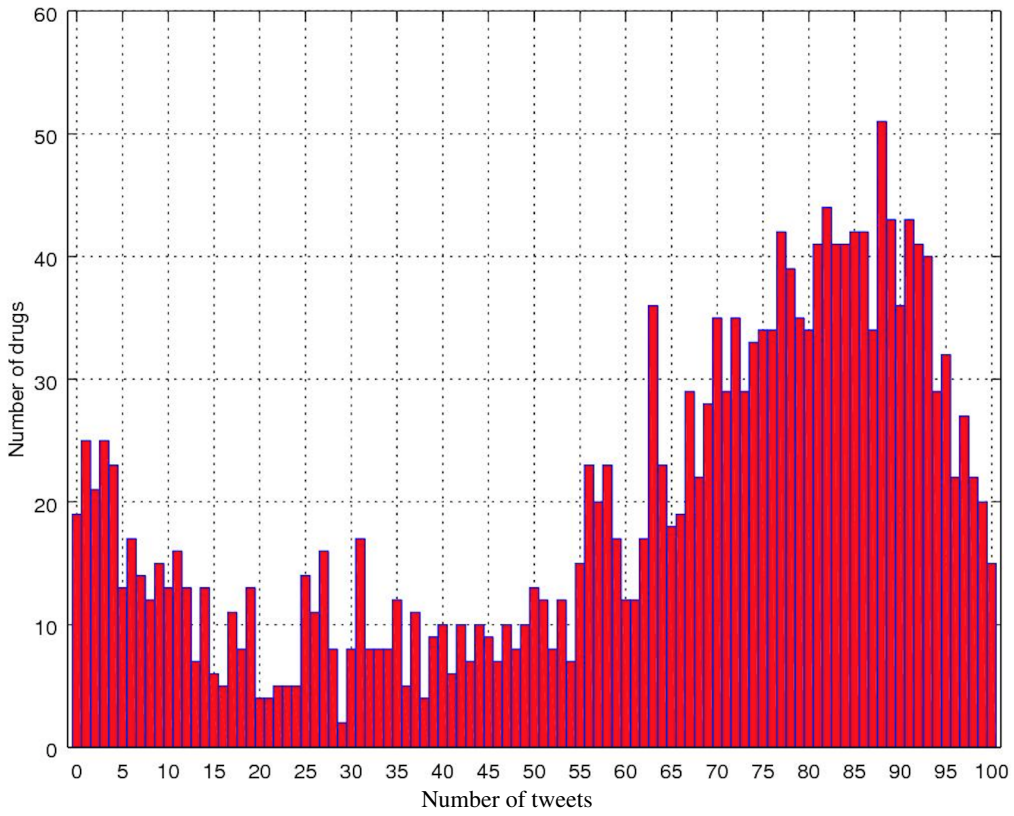

Figure 1: Number of drugs (y) that have a given number of tweets (x). E.g. there are 19 drugs with 0 tweets, and 51 drugs with 88 tweets.

achieved an average F-Score of 0.792, and Accuracy of 0.753. The best three were, by a considerable margin: Emoticons, Panas-t and SANN SA, across all three variants of our corpus. Very briefly, these use the following approaches.

Emoticons: Given lists of positive and neutral emotions, if a text contains an emoticon it is assigned the polarity of the first one, else a text is deemed neutral (Gonçalves et al., 2013a).

Panas-t: Provides association strengths between words and eleven moods including surprise, fear, guilt, joviality, attentiveness, fear, etc. (Gonçalves et al., 2013b).

SANN SA: A dictionary-based rule-based sentiment classifier using the MPQA polarity lexicon (Pappas et al., 2013).

Results for these three classifiers and the majority vote by all 18 systems in ifeel are shown in Table 2; we are only showing results for the raw version of the corpus (see end of Section 4), because results were identical over the three corpus versions, i.e. cleaning, normalising and parsing hashtags did not help with sentiment analysis.

	Emoticons	Panas-t	SANN	Majority vote: (all 18)
F1-measure	**0.941**	0.929	0.920	0.665
Accuracy	**0.936**	0.932	0.932	0.576

Table 2: Tweet-level sentiment analysis results for 3 best sentiment analysis tools on test set B of 300 manually annotated tweets (weighted-average F1-score over POS, NEG and NEU labels; Accuracy).

Out of the 300 tweets in test set B, 259 were annotated as neutral, 22 as positive, and 19 as negative. This gives a very high most-frequent-label baseline of 0.863 Accuracy which the three best classifiers, however, outperform comfortably.

7 Detection of Drug Effect Mentions

We additionally annotated test set B for drug effects with three mention labels: mention of an effect, mention of a property (without specified effect), and none. 81 of the 300 tweets contained an effect mention (sometimes more than one), 10

tweets a property mention, 4 tweets both an effect and a property mention, and 213 contained no mentions.

The experiments in this section were aimed at distinguishing tweets that mention drug effects and properties from tweets that do not. As a starting point we process each of our tweets with MetaMap (Aronson, 2001), a tool for recognising UMLS[9] concepts in text. For a given input text (tweet in our case) MetaMap produces a set of concept vectors as the output, each scored for relevance.

One of the elements in concept vectors indicates 'semantic type' (semtype); semtypes are a set of broad subject categories over all concepts represented in the UMLS metathesaurus. There are 133 different semtypes which fall into six major groupings corresponding to concepts related to organism, anatomical structure, biological function, chemical object, physical object, and idea/concept. For example, a semtype much more commonly seen with tweets that do contain an effect mention is 'Pathologic Function' (83%); examples of semtypes much more commonly seen with tweets that do not contain an effect mention include 'Plant' (96%), 'Receptor' (95%), and 'Laboratory Procedure' (93%).

In this first experiment aimed at detecting effect mentions, we use all 133 semtypes, resulting in binary feature vectors of length 133. We used these paired with the corresponding effect/property labels (+ME = mention of an effect/property, and -ME = none) as training data to train classifiers for this task, using 10-fold cross-validation in testing. The most-frequent-label (-ME) baseline for this task is 0.71 Accuracy.

Table 3 shows results for the three classifiers we tested (Multinomial Naive Bayes, SVM, and Logistic Regression) in terms of overall as well as per-class Recall, Precision and F1-Score. For all classes, SVM performed best with a weighted average F1-Score of 0.873. For the +ME class, the logistic regression classifier did best, with an F1-Score of 0.914. On the -ME class, SVM was best with F1=0.823. Results in Table 3 are for the raw version of the corpus (see end of Section 4); results were identical for the raw and cleaned versions, while additionally parsing hashtags worsened results very slightly for the NB classifier only.

Looking at the Precision scores for +ME, which is arguably the most important measure from the perspective of incorporating information into knowledge graphs, all three classifiers performed extremely well. Especially considering we had a tiny training set, this indicates that the MetaMap semtypes form a highly reliable basis for identifying tweets that mention drug effects/properties.

8 Adverse vs. Beneficial Effects

In test set B, we also annotated those tweets with mentions of an effect or property with a further label encoding whether the effect/property was adverse, beneficial or neutral. 23 tweets were labeled as having an effect/property that is adverse, 57 tweets having a beneficial one, and 11 a neutral one, with 5 tweets having more than one of these.

In the experiments in this section, for the subset of tweets which were identified as containing a drug effect/property mention, we wanted to see whether any of the sentiment analysers would be able to predict whether the effect/property mentioned was an adverse, beneficial or neutral one. We tested the same 18 sentiment analysers against the effect sentiment labels (setting adverse=negative, beneficial=positive, and neutral). Some of the sentiment analysers, despite not being designed for this task, did reasonably well at it; the four best ones were the following:

EmoticonsDS: Uses a large sentiment-scored word list based on the co-occurrence of each token with emoticons in a corpus of over 1.5 billion messages (Hannak et al., 2012).

SenticNet: A semantic and affective resource for concept-level sentiment analysis, modelling the polarities of common-sense concepts and relations between them (Cambria et al., 2014).

SentiWordNet: Lexical SA tool based on WordNet using polarity scores associated with WordNet synsets (Baccianella et al., 2010).

AFINN: Twitter based sentiment lexicon providing emotion ratings for words (Nielsen, 2011).

Table 4 shows the results for these four tools. Interestingly, there is no overlap with the three tools that did best at the standard SA task (Table 2), in fact those three methods were the three *worst* ones at this task.

[9]The Unified Medical Language System (UMLS) "integrates and distributes key terminology, classification and coding standards, and associated resources" (https://www.nlm.nih.gov/research/umls/).

Classifier	Both classes			+ME class			-ME class		
	R	P	F1	R	P	F1	R	P	F1
Multinomial NB	0.847	0.852	0.849	0.848	0.981	0.909	0.961	0.688	0.794
Logistic Regression	0.85	0.851	0.850	**0.855**	0.984	**0.914**	0.969	0.689	0.799
SVM	**0.855**	**0.892**	**0.873**	0.8	**1.0**	0.887	**1.0**	0.71	**0.823**

Table 3: Results for detection of effect mentions (ME) on test set B of 300 manually annotated tweets with 10-fold cross-validation (Recall, Precision, F1-scores for all classes, ME class, and not ME class).

	EmoticonsDS	SenticNet	SentiWordNet	AFINN	Majority vote (all 18)
F1	**0.638**	0.621	0.616	0.615	0.627
Acc	0.592	0.595	0.592	0.597	**0.604**

Table 4: Effect-level sentiment analysis results for 4 best sentiment analysis tools on test set B of 300 manually annotated tweets (weighted-average F1-score over ADV, BEN and NEU labels; Accuracy).

9 Conclusions and Further Work

In this paper we described our new corpus of 125,669 tweets for 1,999 drug names. The corpus includes one randomly selected subset (A) of 300 tweets which is annotated for language, and another set (B) of 300 random tweets which is annotated for overall sentiment and drug effect mentions. The corpus represents many more drug names than in comparable existing corpora, but so far has only a small number of annotated tweets.

We reported results for language identification and sentiment analysis. One of the language identification tools tested (langid.py) outperforms the Twitter language tags in this domain. Tweet-level sentiment analysis achieved a best result of 0.94 weighted average F1-Score (Emoticons method).

Perhaps the most surprising result we report is that a straightforward approach to training a classifier to distinguish tweets that mention drug effects from those that do not, achieves an overall F1-Score of 0.873 and perfect precision for the positive class (+ME). This indicates that MetaMap semtypes are a strong basis for predicting drug effect mentions. However, using sentiment analysis tools for the new task of predicting the polarity of an effect (adverse vs. beneficial vs. neutral) achieved best results of just 0.64 weighted average F1-Score (EmoticonsDS).

In future work, we are planning to expand our annotations to allow more effective training of classifiers in particular to address the latter task, as well as generally to expand our corpus to include tweets containing common, colloquial and generic names and common misspellings of our drug names. In terms of methodology, we are currently working on drug effect extraction, i.e. identifying the span of words in a tweet that describes the effect.

References

M Abou Taam, C Rossard, L Cantaloube, N Bouscaren, G Roche, L Pochard, F Montastruc, A Herxheimer, JL Montastruc, and H Bagheri. 2014. Analysis of patients' narratives posted on social media websites on Benfluorex's (mediator®) withdrawal in france. *Journal of Clinical Pharmacy and Therapeutics*, 39(1):53–55.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Jiang Bian, Umit Topaloglu, and Fan Yu. 2012. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international Workshop on Smart Health and Wellbeing*, pages 25–32. ACM.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*, pages 1515–1521. AAAI Press.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen OConnor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. 2014. 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*.

Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. 2013a. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38. ACM.

Pollyanna Gonçalves, Fabrício Benevenuto, and Meeyoung Cha. 2013b. Panas-t: A pychometric scale for measuring sentiments on twitter. *arXiv preprint arXiv:1308.1857*.

Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. 2012. Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *ICWSM*. Citeseer.

J. Lardon, R. Abdellaoui, F. Bellet, H. Asfari, J. Souvignet, N. Texier, M.C. Jaulent, M.N. Beyens, A. Burgun, and C. Bousquet. 2015. Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*, 17:171.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 17–25.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Filipe Nunes Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Fabrício Benevenuto, and Marcos André Gonçalves. 2010. A benchmark comparison of state-of-the-practice sentiment analysis methods. *ACM Transactions on Embedded Computing Systems*, 9(4).

Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos. 2013. Distinguishing the popularity between topics: A system for up-to-date opinion retrieval and mining in the web. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 197–209. Springer.

Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

WHO. No date. World Health Organization English Glossary. Online document, http://www.who.int/medicines/areas/coordination/English_Glossary.pdf. Accessed: 27 July 2016.