READ-BIOMED-SS: ADVERSE DRUG REACTION CLASSIFICATION OF MICROBLOGS USING EMOTIONAL AND CONCEPTUAL ENRICHMENT

BAHADORREZA OFOGHI[1], SAMIN SIDDIQUI[1], and KARIN VERSPOOR[1,2]

[1]Department of Computing and Information Systems
[2]Health and Biomedical Informatics Centre
The University of Melbourne
Melbourne, Victoria, Australia

{bahadorreza.ofoghi,karin.verspoor}@unimelb.edu.au

Abstract

This paper describes the READ-BioMed team's participation in the Social Media Mining Shared Task of the Pacific Symposium on Biocomputing 2016. The task had a focus on Adverse Drug Reaction classification of Twitter microblogs. Our READ-BioMed Social media Surveillance system (READ-BioMed-SS) implemented a few lexical normalization processes and employed existing tools to enrich tweet texts before applying a machine learning-based classifier on the tweets. The conceptual enrichment of tweets was based on the sentiment of the tweets, emotion classes, some UMLS Metathesaurus concepts, as well as drug, chemical substance, and disease mentions. The best performance of READ-BioMed-SS on the official test set was achieved after a bag-of-words representation for tweets was enriched with sentiment analysis, emotion classes, and specific UMLS Metathesaurus concepts.

Keywords: Adverse drug reaction, UMLS, emotion, sentiment, SMOTE

1 Introduction

Members of the general public may use social media to share their health experiences with others, for feedback or simply to vent their concerns. Social media therefore provides a rich potential resource for monitoring public health issues. For drugs, this specifically means monitoring of post-market, spontaneous reporting of adverse events linked to the drugs [6]. The PSB 2016 Social Media Mining (SMM) shared task aims to explore the application of natural language processing techniques to microblog data for the purpose of surveillance of adverse drug reactions (ADRs). The READ-BioMed team's submission to the SMM shared task addressed Task 1, "Binary classification of ADRs", which involves classifying a tweet from Twitter as containing an ADR, or not.

2 Background

There have been several prior approaches to binary classification of drug-related adverse reactions on social media [2, 5, 9]. Challenges that exist in surveillance of social media include the language gap between medical and patient vocabulary, noise and biases in the information, and potential lack of specificity to individual drugs [6]. To address this gap, many systems take advantage of lexicons that capture terms and synonyms for a range of relevant concepts, including drug names, clinical symptoms, and specifically "adverse reaction" terms, coupled with machine learning methods [11, 6].

Some recent ADR classification systems have also augmented named entity recognition of drugs, diseases, and symptoms with other natural language processing features (e.g., n-grams and topic models) to improve ADR performance [11].

In our work, we extend the existing strategies by further pre-processing of texts to assess both the general sentiment of the tweet (positive/neutral/negative) and the finer-grained emotional stance of the tweet (anger/sadness/happiness/etc.). We find that incorporating this information into the tweet representation improves the performance of the adverse drug reaction classification.

3 READ-BioMed ADR Classification Approach, PSB SMM 2016

The READ-BioMed social surveillance team submission to the SMM Task 1 utilised supervised machine learning techniques to identify Twitter microblogs that contained ADR mentions or relevant meaning. To train the classification system, we used the training data provided by the task organizers. A number of preprocessing and normalization steps were applied, as will be discussed in the next sections. The tweets were then enriched with several types of features that were extracted using other machine learning-based tools and annotation systems. This resulted in a number of different experimental set-ups, with varying features. For each training scenario, a 10-fold cross validation procedure with a SVM classifier was carried out to estimate the effectiveness of the preprocessing and feature engineering steps.

Once the test set of the task was released, the READ-BioMed Social media Surveillance system (READ-BioMed-SS) was run over the test set with these different preprocessing and feature engineering settings. Then, a selected set of five runs were reported to the organizers.

3.1 Data preprocessing

In the first step, we retrieved the training set of tweets, which included 5,329 tweets in total, 603 labelled as positive or ADR and 4,726 tweets as negative or non-ADR. The next step was preprocessing and normalization of the data. All URLs, email addresses, mentions (i.e., @replies and @usernames), and hash tags were replaced by "url", "emailAddress", "atSign", and "hashTag", respectively. In addition, any mention of drug dosages (with "mg" or "MG" postfixes) were replaced with "drgOrMedDsg". Stop-words were also removed from the tweets and the texts were converted to lowercase. In the last step, for some scenarios, the Stanford lemmatizer was used to replace all tokens with their lemmas.

3.2 Feature enrichment

After the tweets were preprocessed and normalized, they were utilized in different ways to train a number of READ-BioMed-SS binary ADR classifiers. In their basic form, the tweets were modelled as bag-of-words features only without any further analysis. Consequently, five different types of additional features were used to enrich the representation of the tweets:

- Sentiments: Stanford Sentiment Analyzer [12] was used to find the sentiment of the tweets (i.e., negative, neutral, or positive). This will represent the overall subjective information of the tweets.
- Emotion Classes: A previously trained Emotion Classifier [10] was utilized to find the emotion that is expressed in a tweet. The classifier labelled each tweet with one of the nine classes "anger", "disgust", "happiness", "sadness", "surprise", "fear", "news-related", and "criticism". The emotion classifier makes use of a number of emotion vocabularies, emoticons, and sentiments to find the overall emotion of a tweet.
- UMLS mentions: The MetaMap [1] Web API was used to find some of the UMLS Metathesaurus concepts in the tweet texts. The UMLS Semantic Types that we considered included "Disease or Syndrome", "Clinical Drug", "Injury or Poisoning", and "Pharmacologic Substance". We believe these are the types of mentions that can be present in ADR related sentences.
- Chemicals/drugs: tmChem [8] was also utilised to find specific mentions of any drugs and chemicals in the texts of the tweets.
- Diseases: BANNER [7] with the disease model was used to find disease mentions that may not have been captured using MetaMap. Some ADRs may result in conditions also coded as disease, e.g. skin rash.

Once each of the above features was extracted for a tweet, they were added to the core bag-of-words representation of the tweet. This resulted in five different configurations for the training (and test) data features, starting with the basic bag-of-words features and incrementally adding sentiments, emotion classes, the UMLS concepts, and chemicals/drugs/diseases.

4 System Training

A READ-BioMed-SS classifier was trained with 20 different feature configurations for the training data. The first 5 configurations included tweets only (bag-of-words with no extra features), as well as the tweets incrementally enriched with each of the above-mentioned features (with each feature type added monotonically to the previous features, in the order listed above, i.e. sentiments, then sentiments & emotions, sentiments & emotions & UMLS, etc.). Then, we trained the system with both lemmatized and non-lemmatized versions of the training data (5 configurations for each version, resulting in a total of 10 configurations).

Since the original training data is imbalanced (603 ADR-labelled tweets out of 5,329), an over-sampling technique was used to create a balanced data set. For this, several runs of Synthetic Minority Oversampling TEchnique (SMOTE) [3] was utilised, resulting in a data set with 9,550 tweets, 4,824 ADR and 4,726 non-ADR-labelled instances. This data set was used with the 10 previously defined feature configurations, to create 10 more classifiers.

Table 1: READ-BioMed-SS 10-fold cross validation results on the training configurations where tweet tokens were lemmatized. Note: bow=bag-of-words, sent=sentiment, ec=emotion class, dcd=drug/chemical/disease.

Data set	Features	non-ADR			ADR		
		precision	recall	F1	precision	recall	F1
original	bow	0.927	0.958	0.942	0.552	0.406	0.468
	bow+sent	0.928	0.957	0.942	0.555	0.418	0.477
	bow+sent+ec	0.928	0.957	0.942	0.555	0.418	0.477
	bow+sent+ec+umls	0.930	0.955	0.942	0.555	0.436	0.488
	bow+sent+ec+umls+dcd	0.929	0.956	0.943	0.557	0.431	0.486
original+smote	bow	0.928	0.956	0.942	0.956	0.927	0.941
	bow+sent	0.928	0.956	0.942	0.955	0.927	0.941
	bow+sent+ec	0.929	0.958	0.943	0.958	0.928	0.943
	bow+sent+ec+umls	0.928	0.955	0.941	0.955	0.927	0.941
	bow+sent+ec+umls+dcd	0.927	0.955	0.941	0.955	0.926	0.940

To evaluate the effectiveness of each preprocessing, normalization, and feature engineering procedure, a 10-fold cross validation process was utilized with SVM classifier. WEKA [4] machine learning toolkit was used for the experiments. The results of this analysis is summarized in Table 1. For simplicity, only the results using lemmatization are shown; the results without lemmatization are similar.

As shown in Table 1, over-sampling of the ADR-labelled instances resulted in significant classification improvements over classifiers trained with original imbalanced data set, especially in terms of precision, recall, and F1 measure on the high-importance ADR class, which has a small number of tweets in the original training set.

5 System Testing

After this training and development phase, to obtain the results on the released test data set, READ-BioMed-SS classifiers were trained with the full training data, for each feature configuration. Each classifier was then applied to the test data, with a matched feature representation. Since we had 20 classifiers, 20 prediction configurations were produced for the test data. The results obtained on each test configuration (considering lemmatization only) are summarized in Table 2. Since we do not have access to the gold standard test set (i.e., no class labels), we can only count the instances labelled with each class.

Table 2: READ-BioMed-SS results on the test data, using a range of different features, where tweet tokens were lemmatized. Note: TS=training set, bow=bag-of-words, sent=sentiment, ec=emotion class, dcd=drug/chemical/disease.

Features	TS=Original		TS=Original+smote	
	#non-ADR	#ADR	#non-ADR	#ADR
bow	4546	349	4523	372
bow+sent	4547	348	4516	379
bow+sent+ec	4515	380	4500	395
bow+sent+ec+umls	4486	409	4475	421
bow+sent+ec+umls+dcd	4507	388	4496	399

6 System Selection and Official Results

For the Social Media Shared Task on ADR classification, the READ-BioMed team submitted 5 system runs on the test data. Our team had to decide the results of which system runs, from the set of 20 runs mentioned in previous sections, to submit. Our decision was based on the performance of the classifiers on the training data, as well as statistical analysis of the classification outputs of each system run. We decided to submit the results of the two runs with lemmatization and the feature sets including bag-of-words, sentiments, emotion classes, and the UMLS concepts when the system was trained with both original and over-sampled training configurations. These were the two systems that mostly resulted in the largest numbers of ADR-labelled instances. We also decided that the basic bag-of-words runs of each data set should be reported to understand whether the addition of extra features has been practically useful. To select the final run for submission, we carried out a Kappa statistical agreement analysis between the output labels of the systems trained with the improved balanced training set and selected the run that had the largest difference (i.e., the smallest agreement rate) with the runs already selected from this set of system runs. Table 3 summarizes the settings of each submitted run as well as the official results received from the task organizers for each run.

Table 3: READ-BioMed-SS official task results. Note: TS=training set, Prec=ADR precision, Rec=ADR recall, F1=ADR F-measure, Acc=accuracy, bow=bag-of-words, lem=lemmatization, sent=sentiment, ec=emotion class, dcd=drug/chemical/disease. Rank shows the rank of the run in the set of five system runs.

System run	Rank	TS/Features/Lemmatization	Prec	Rec	F1	Acc
READ-Biomed-010-1a	4	orgiginal/bow/+lem	0.358	0.332	0.344	0.903
READ-Biomed-010-1b	2	orgiginal/bow+sent+ec+umls/+lem	0.342	0.371	0.356	0.897
READ-Biomed-010-1c	3	orgiginal+smote/bow/+lem	0.357	0.353	0.355	0.901
READ-Biomed-010-1d	1	orgiginal+smote/bow+sent+ec+umls/+lem	0.340	0.379	0.358	0.895
READ-Biomed-010-1e	5	orgiginal+smote/bow+sent+ec+umls+dcd/-lem	0.312	0.326	0.319	0.893

7 Discussion and Concluding Remarks

From our 5 system submissions, READ-Biomed-010-1d performed the best (ADR F1=0.358) according to the official evaluations (see Table 3 for more details). This was the system that was trained with the improved over-sampled training set and utilised the lemmatized bag-of-words features in conjunction with extracted sentiments, emotion classes, and the UMLS concepts. The system run with the lowest performance (ADR F1=0.319) was READ-Biomed-010-1e which did not lemmatize tokens however utilised all the enrichment techniques (sentiments, emotion classes, UMLS concepts, and drug/chemical/disease mentions). Although from the results it is not directly possible to understand why READ-Biomed-010-1e performs worse than READ-Biomed-010-1d (the former uses drug/chemical/disease mentions but does not lemmatize tokens), referring to the results of System Testing in Table 2, we find that READ-Biomed-010-1d returned 421 ADR-labelled tweets versus READ-Biomed-010-1e's 399 ADR-labelled instances. This suggests that the addition of extra drug/chemical/disease mentions did not positively contribute to the ADR classification system.

Addition of emotion classes has also improved our official system runs, as evidenced by the fact that READ-Biomed-010-1b and READ-Biomed-010-1d have outperformed READ-Biomed-010-1a and READ-Biomed-010-1c. This finding is consistent with those in the System Training and System Testing phases especially when tokens were lemmatized (see Table 1 and Table 2).

Another aspect of the official results was the fact that the systems that were trained with the over-sampled data (i.e., the last three runs in Table 3) did not reach significantly higher performances compared with the two system runs that were only trained on the original imbalanced data set. This is in contrast to the substantial improvements achieved in the system training phase when using SMOTE over-sampling for the minority class of ADR-labelled tweets (see Table 1).

References

[1] A.R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceedings of the 2001 AMIA Annual Symposium*, pages 17–21, Washington, DC, 2001.

[2] J. Bian, U. Topaloglu, and F. Yu. 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, 2012.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[5] K. Jiang and Y. Zheng. Mining twitter data for potential drug effects. In Hiroshi Motoda, Zhaohui Wu, Longbing Cao, Osmar Zaiane, Min Yao, and Wei Wang, editors, *Advanced Data Mining and Applications*, volume 8346 of *Lecture Notes in Computer Science*, pages 434–443. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-53913-8.

[6] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris. Text and data mining techniques in adverse drug reaction detection. *ACM Comput. Surv.*, 47(4):56:1–56:39, May 2015. ISSN 0360-0300. doi: 10.1145/2719920.

[7] R. Leaman and G. Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, pages 13:652–663, 2008.

[8] R. Leaman, C-H Wei, and Z. Lu. tmChem: A high performance tool for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7, 2015.

[9] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, page ocu041, 2015.

[10] B. Ofoghi, M. Mann, and K. Verspoor. Towards early discovery of salient health threats: A social media emotion classification technique. In *Proceedings of Pacific Symposium on Biocomputing (PSB)*, Hawaii, 2016.

[11] A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196–207, 2015.

[12] R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, USA, 2013.