ADVERSE DRUG REACTION DETECTION USING AN ADAPTED SENTIMENT CLASSIFIER*

DOMINIC EGGER
Zurich University of Applied Science
Winterthur 8400 Switzerland
eggo@zhaw.ch

FATIH UZDILLI
Zurich University of Applied Science
Winterthur 8400 Switzerland
uzdi@zhaw.ch

MARK CIELIEBAK
Zurich University of Applied Science
Winterthur 8400 Switzerland
ciel@zhaw.ch

We describe a classifier to determine whether a tweet message mentions an adverse drug reaction (ADR). The classifier produced for this task is based on a system originally intended for sentiment analysis, with additional specific features and resources for ADR detection. The experiments show that the additional features have not produced a significant increase in F-score. However, the clear increase in score by using sentiment specific features on top of general text analysis features shows that there is a correlation between ADR and sentiment analysis.

1. Introduction

In the large amount of social media posts on platforms like Twitter there are many messages discussing medications. Being able to automatically detect messages containing mentions of adverse drug reactions (ADR) would be in the interest of public health and government agencies. They would be able to aggregate more information over a wider group of people about drugs and their side effects.

To increase the quality of research in this area, the DIEGO lab of Arizona State University organized a shared task. The goal was to create a system able to detect ADR mentions in tweets. The objective was the highest possible F1-score on the ADR class on a hidden test set. A labeled training set[14,2] containing 806 ADR tweets and 6761 Non-ADR tweets was provided by the organizers. Our submission ranked 5th out of 8 participants with an F-Score of 0.3174. The best submission had an F-score of 0.4195.

* This project was funded by the CTI 'Commission for Technology and Innovation'

2. Description of our Approach

Our system is based on our existing sentiment classification system.[4,15] The system is primarily based on textual features used by a Support Vector Machine classifier. One assumption tested during the experiments is the existence of a correlation between sentiment and ADR tasks. Also, our work about the sentiment classifier has shown that the general text features without sentiment specific information still performs satisfactory for the sentiment task. We expected the same for ADR.

Our experiments showed that our initial assumption holds for the training data (see in section 'Results'). Therefore we used the whole feature set of our sentiment system and enriched it with ADR specific resources, such as lexica using distant supervised labels. We used a modified version of LibLinear[1] to generate our classifiers. They were trained using the L1-regularized L2-loss support vector classification algorithm. To deal with the unbalanced nature of the dataset, we trained several sub-classifiers, each with all positive ADR training examples and pairwise disjoint subsets of the negative examples. The final classification of a new tweet was done by majority vote.

3. Features

All features were generated after tokenizing the tweets using ArkTweetNLP[11] and a preprocessing step where URLs and usernames were normalized. The negation scope of the tweet was marked for each single token as in.[13] Smaller features not listed explicitly are: The number of hashtags used in the tweet. The number of negated contexts found in a tweet, as detected by Pant *et. al.*[13] The Number of elongated tokens found in a tweet. The number of tokens entirely written in capital letters and the number of occurrence for each POS tag found in a tweet

- **n-Grams:** n-Grams of tokens of the text for n = 1..4. Additionally, for n = 3...5 we used a wildcarding technique where we replaced the one or more tokens in every combination with POS tags, lemmas or wildcards. As well as n-grams for characters for n=3..6.
- **Clustering:** For each word cluster, we decided whether it contains a word from the tweet. For this we used the Twitter word clusters[12] and word clusters from Nikfarjam *et al.*[10]
- **Word Embeddings:** Part of our features was the Max, Min and mean values of word embedding vectors. The source of Word Embeddings we used is Nikfarjam *et al*.[10]
- **Lexica Features:** We used a wide variety of lexica that can broadly be put into two categories. The sentiment lexica have usually some sort of scoring associated with each token. With these lexica we extracted the total summed score per tweet for each sentiment class (negative, positive, and neutral) as well as the score for each class, specifically for the last token in the tweet. We used the following sentiment lexica: NRC Canada,[9] Bing Liu's Sentiment Words,[3,7] Sen140 Lexicon[5,17] and MPQA Lexicon.[16] The other category of lexica we used were ADR specific ones. For those lexica that provide some sort of scoring, we utilized the same set of features as described above.

For those lexica not providing a scoring we used a simple counter of tokens that could be matched to those lexica. We generated some of these lexica ourselves, which are further described in section titled 'Lexica Generation'. Others were already available online, such as a lexicon by the Arizona State University[10] and the SIDERS package.[6]

4. Lexica Generation

Based on the approach described in,[8] we created a list of 4 million English tweets containing drug names and symptoms from the SIDERS package. We used the distant supervised labels 'contains drug mention', 'contains symptom mention', 'contains both' and 'contains none'. We produced lexica of tokens with its strength of association score for each of these labels. Before we calculated the scoring, we removed the drug name or symptom name as we were aiming to get context information.

5. Results

We achieved an F1-score of 0.3175 on the final test data. Because the test data was not available at the time of writing, we provide the 10-fold cross validation results for each subclassifier in Table 1. Interestingly, the cross validation scores are far better than the final result in the shared task. The reason is currently unknown and part of a future investigation as soon as the test set becomes available. However, Table 1 shows that the sentiment resources increased the score. On the other hand, the ADR specific resources did not show any improvement.

Table 1: 10-fold F1 cross validation scores on the ADR Class of the subclassifiers

Classifier	None	Sentiment	ADR	Both
1	0.7124	0.7400	0.7120	0.7261
2	0.7258	0.7504	0.7250	0.7641
3	0.7189	0.7338	0.7028	0.7248
4	0.7026	0.7271	0.7112	0.7431
5	0.6983	0.7344	0.7125	0.7436
6	0.7172	0.7538	0.7119	0.7350
7	0.7228	0.7247	0.6934	0.7288
Median/Mean	0.**7172**/0.**7140**	0.**7344**/0.**7378**	0.**7119**/0.**7098**	0.**7350**/0.**7379**

6. Conclusion

Our initial assumption was that sentiment resources are useful for ADR specific tasks. Our experiments show that this holds true, at least for the actual case. On the other hand, usage of our ADR resources did not perform as hoped. The reason might be that our strong sentiment lexica are very powerful and proven to work on various sentiment data sets. In

contrast to that, the ADR lexicon was our first attempt for an ADR task. Also, the distant labels for sentiment are quite intuitive while our distant labels for ADR may be just the wrong direction. Future work needs to investigate this further.

References

1. Fan, Rong-En et al. "LIBLINEAR: A library for large linear classification." The Journal of Machine Learning Research 9 (2008): 1871-1874.
2. Ginn et al., 2014. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. BIOTXTM.
3. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining 22 Aug. 2004: 168-177.
4. Jaggi, Martin, Fatih Uzdilli, and Mark Cieliebak. "Swiss-Chocolate: Sentiment Detection using Sparse SVMs and Part-Of-Speech n-Grams." SemEval 2014 (2014): 601.
5. Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad. "Sentiment analysis of short informal texts." Journal of Artificial Intelligence Research (2014): 723-762.
6. Kuhn, Michael et al. "A side effect resource to capture phenotypic effects of drugs." Molecular systems biology 6.1 (2010): 343.
7. Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." Proceedings of the 14th international conference on World Wide Web 10 May. 2005: 342-351.
8. Mohammad, Saif M, and Svetlana Kiritchenko. "Using hashtags to capture fine emotion categories from tweets." Computational Intelligence 31.2 (2015): 301-326.
9. Mohammad, Saif M, Svetlana Kiritchenko, and Xiaodan Zhu. "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets." Second Joint Conference on Lexical and Computational Semantics (* SEM) 14 Jun. 2013: 321-327.
10. Nikfarjam, Azadeh et al. "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features." Journal of the American Medical Informatics Association (2015): ocu041.
11. Owoputi, Olutobi et al. "Part-of-speech tagging for Twitter: Word clusters and other advances." School of Computer Science, Carnegie Mellon University, Tech. Rep (2012).
12. Owoputi, Olutobi et al. "Improved part-of-speech tagging for online conversational text with word clusters." 2013: 380.
13. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 6 Jul. 2002: 79-86.
14. Sarker and Gonzalez, 2014. Protable automatic text classificaiton for adverse drug reaction detection via multi-corpus training. Journal of Biomedical Informatics.
15. Uzdilli, Fatih et al. "Swiss-Chocolate: Combining Flipout Regularization and Random Forests with Artificially Built Subsystems to Boost Text-Classification for Sentiment."

16. Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." Proceedings of the conference on human language technology and empirical methods in natural language processing 6 Oct. 2005: 347-354.
17. Zhu, Xiaodan, Svetlana Kiritchenko, and Saif M Mohammad. "Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets." SemEval 2014 443 (2014).