NTTMU System in the 2nd Social Media Mining for Health Applications Shared Task

Chen-Kai Wang[1], Nai-Wun Chang[2], Emily Chia-Yu Su, PhD[1], Hong-Jie Dai, PhD[3]
[1]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan, R.O.C.; [2]Institute of Information Science, Academia Sinica, Taipei, Taiwan; [3]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan, R.O.C.

Abstract

In this study, we describe our methods to automatically classify Twitter posts describing events of adverse drug reaction and medication intake. We developed classifiers using linear support vector machines (SVM) and Naïve Bayes Multinomial (NBM) models. We extracted features to develop our models and conducted experiments to examine their effectiveness as part of our participation in AMIA 2017 Social Media Mining for Health Applications shared task. For both tasks, the best-performed models on the test sets were trained by using NBM with n-gram, part-of-speech and lexicon features, which achieved F-scores of 0.295 and 0.615, respectively.

Introduction

We participated in two tasks of the 2nd Social Media Mining for Health Applications Shared Task. One is the automatic classification of adverse drug reaction mentioning posts (Task 1), which focuses on classifying tweets mentioned adverse drug reactions for pharmacovigilance [1]. The other is the task of automatic classification of posts describing medication intake (Task 2). The goal of the task 2 is to automatically classify tweets posted by users describing drug intake of the paste, hoping to improve the accuracy to monitor the safety of drugs [2].

Methods

We formulated both tasks as classification problems and used support vector machine (SVM) and naïve Bayes multinomial (NBM) to develop our classifiers. For task 1, the organizers provided a training set and a development set with 8,029 and 3,549 tweets, respectively. Moreover, the binary annotation indicating the presence or absence of ADRs for each tweet. We used this training and development set to engineer features and develop our classifiers. For the task 2, 7,463 and 2,107 tweets are provided as a training set and a development set. And for each tweet, three classes are assigned such as (1) personal medication intake, tweets in which the user clearly expresses a personal medication intake/consumption; (2) possible medication intake, tweets that are ambiguous but suggest that the user may have taken the medication; and (3) non-intake, tweets that mention medication names but do not indicate personal intake. The classifiers were evaluated using a 10-fold cross validation (CV) on both datasets with precision, recall and F-measure metrics.

Initially, we pre-processed the datasets to remove Twitter specific characters like hashtags, usernames, and repetition of certain alphabets. In addition, we used regular expressions to detect mentions of dosage and replaced them with the "@DSG" symbol. For instance, "1.1 mg" will be replaced to "@DSG". This was followed by stemming using the Snowball stemmer[1] and tokenizing using the tokenizer developed by Owoputi, O'Connor, Dyer, et al. [3]. Twitter posts are very short and in order to preserve the information expressed, we did not remove any stop words.

After the pre-processing, we extracted various features to train our models. The SVM with a linear kernel was trained with the sequential minimal optimization algorithm. We used the implementations provided by Weka [4] to develop our systems. The features we used included:

- *N*-gram features: The *n*-gram features in which the range of *n* was set to one to three, including unigram, bigram and trigram.
- Part-of-speech (POS) tags: The POS tags generated by the Twitter NLP tool[2].

[1] http://snowball.tartarus.org/
[2] http://www.cs.cmu.edu/~ark/TweetNLP/

- Lexicon-based features: We used the ADR lexicon compiled in our previous work [5] to mark their presence and developed two binary features for a tweet; one is the presence of drug names and the other is presence of ADR mentions.

In addition to the above features, we have tried to exploit a likely positive dataset [6] and employed different term weighting methods, such as the transformed weight-normalized complement Naïve Bayes (TWCNB) [7]. Naïve Bayes classifier and the weighted features, such as term frequency, inverse document frequency, length normalization and complement class weighting, are used as the factors for TWCNB. Unfortunately, we could not achieve any significant improvement over the above feature sets. We will report the details in the Results section.

Results

Table 1 and 2 show the results of the 10-fold CV on the training sets of the task 1 and 2 respectively. The standard precision (P), recall (R) and F-measure (F) are used to report the performance. Configuration 1 and 2 show the performance of baseline models trained with *n*-gram features. The F-measure are individually treated as the baseline scores used for calculating the last column of Table 1 and 2 for SVM and NBM. In configuration 3 and 4, we included the POS information as new features. Configuration 5 and 6 show the results after preprocessing the dosage information. Table 1 also shows the results after applying SMOTE (Synthetic Minority Over-sampling Technique) for the class imbalance problem, stop word filtering, spell correction and attribute selection based on information gain. However, we didn't see any improvement on F-measure.

For task 1, the configuration (14) adopting NBM algorithm with all proposed features achieves the highest F-measure, 49.92%, here. And for task 2, the same configuration (denoted as 4 in Table 2) also achieved the highest F-measure, 63.34%.

Table 1. 10 fold CV on the training set of the task 1.

Configuration		P	R	F	Diff of F
(1)	SVM	0.550	0.441	0.490	-
(2)	NBM	0.384	0.635	0.479	-
(3)	(1) + PoS	0.539	0.438	0.483	-0.007
(4)	(2) + PoS	0.398	0.659	0.496	+0.017
(5)	(3) + DSG	0.541	0.435	0.482	-0.008
(6)	(4) + DSG	0.482	0.662	0.496	+0.017
(7)	(5) + SMOTE	0.514	0.464	0.488	-0.002
(8)	(6) + SMOTE	0.363	0.640	0.463	-0.016
(9)	(5) + Stop Word	0.539	0.401	0.460	-0.030
(10)	(6) + Stop Word	0.395	0.661	0.495	+0.016
(11)	(5) + Spell	0.544	0.443	0.488	-0.002
(12)	(6) + Spell	0.396	0.670	0.498	+0.019
(13)	(11) + Lexicon	0.545	0.443	0.489	-0.001
(14)	(12) + Lexicon	0.397	0.671	**0.499**	**+0.020**

(15)	(13) + Attribute Select	0.692	0.306	0.425	-0.065
(16)	(14) + Attribute Select	0.428	0.588	0.495	+0.016

Table 2. 10 fold CV on the training set of the task 2.

Configuration	P	R	F
(1) SVM + PoS + DSG + Spell	0.622	0.640	0.631
(2) NBM + PoS + DSG + Spell	0.668	0.599	0.631
(3) (1) + Drug	0.621	0.640	0.630
(4) (2) + Drug	0.669	0.601	0.**633**

Table 3 shows the results on the test set. For each task, we submitted three runs. The first and second runs are based on the best configurations observed in our experiments on the training set as shown in Table 1 and 2. For the third run, we employed attribute selection to select features for NBM models.

Table 3. The results on the test set of the task 1 and 2.

Configuration	P	R	F
PSB SMM4H Shared Task 1 Results			
(1) NBM+DSG+Pos+Hun+ Drug	0.213	**0.433**	0.286
(2) SMO+DSG+Pos+Hun+ Drug	**0.362**	0.249	**0.295**
(3) (1) + Attribute Select	0.226	0.403	0.29
PSB SMM4H Shared Task 2 Results			
(1) NBM+DSG+Pos+Hun+ Drug	**0.69**	0.554	0.614
(2) SMO+DSG+Pos+Hun+ Drug	0.644	**0.588**	**0.615**
(3) (1) + Attribute Select	0.662	0.572	0.614

Conclusion

In this paper, we gave a briefly introduction of our systems based on SVM and NBM algorithms and conducted experiments to study the effectiveness of different features and preprocessing. We observed that the best configurations for both tasks were based on the spell-checked and dosage-replaced tweets along with *n*-gram, POS and lexicon features.

References

[1] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *J Biomed Inform*, vol. 53, pp. 196-207, Feb 2015.

[2] A. Klein, A. Sarker, M. Rouhizadeh, K. O'Connor, and G. Gonzalez, "Detecting Personal Medication Intake in Twitter: An Annotated Corpus and Baseline Classification System," *BioNLP 2017*, pp. 136-142, 2017.

[3] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2013.

[4] G. Holmes, A. Donkin, and I. H. Witten, "Weka: A machine learning workbench," in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, 1994, pp. 357-361.

[5] J. Jonnagaddala, T. R. Jue, and H.-J. Dai, "Binary classification of Twitter posts for adverse drug reactions," presented at the Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Big Island, Hawaii, 2016.

[6] R. T.-H. Tsai, H.-C. Hung, H.-J. Dai, Y.-W. Lin, and W.-L. Hsu, "Exploiting Likely-Positive and Unlabeled Data to Improve the Identification of Protein-Protein Interaction Articles," presented at the 6th InCoB - Sixth International Conference on Bioinformatics, 2007.

[7] M. Timonen, "Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion," 2013.