Drug-use Identification from Tweets with Word and Character N-grams

Çağrı Çöltekin
Department of Linguistics
University of Tübingen, Germany
ccoltekin@sfs.uni-tuebingen.de

Taraka Rama
Department of Informatics
University of Oslo, Norway
tarakark@ifi.uio.no

Abstract

This paper describes our systems in social media mining for health applications (SMM4H) shared task. We participated in all four tracks of the shared task using linear models with a combination of character and word n-gram features. We did not use any external data or domain specific information. The resulting systems achieved above-average scores among other participating systems, with F_1-scores of 91.22, 46.8, 42.4, and 85.53 on tasks 1, 2, 3, and 4 respectively.

1 Introduction

The increasing use of social media platforms world wide offers an interesting application of natural language processing tools for monitoring public health and health-related events on the social media. The social media mining for health applications (SMM4H) shared task (Weissenbacher et al., 2018) hosts four tasks aiming to identify mentions of different aspects medication use on Twitter. Briefly, the tasks and their descriptions are:

Task 1: Automatic detection of posts mentioning drug names.
Task 2: Automatic classification of posts describing medication intake.
Task 3: Automatic classification of adverse drug reaction mentioning posts.
Task 4: Automatic detection of posts mentioning vaccination behavior.

All tasks, except Task 2 are binary classification tasks. Task 2 requires three-way classification, including an uncertain class indicating posts mentioning possible medication intake.

For all tasks, we used linear SVM classifiers with character and word bag-of-n-gram features. We also experimented with other methods, including deep learning methods with gated RNNs for building document representations. However, SVM models achieved best results on the development data. As a result, we only submitted results using linear SVMs, and we will only describe and discuss results of these model in this paper.

2 Methods and Experimental Procedure

We use the same general model for all tasks: linear SVM classifiers with character and word bag-of-n-gram features. Tokenization was done using a simple regular expression tokenizer that splits the text into consecutive alphanumeric and non-space, non-alphanumeric tokens. For each text to be classified, we extracted both character and word n-grams of order one up to a certain upper limit (specified below). All features are combined in a flat manner as a single text-feature matrix. We experimented with two feature weighting methods: tf-idf (Jurafsky and Martin, 2009, p.805) and BM25 (Robertson et al., 2009). The weighted features are then used for training an SVM classifier. We used one-vs-rest multi-class strategy when training the SVM classifier for task 2. All models were implemented in Python, using scikit-learn machine learning library (Pedregosa et al., 2011). The models are similar to the models we used in a few other text classification tasks (Çöltekin and Rama, 2018; Rama and Çöltekin, 2017; Çöltekin and Rama, 2017), where the models are explained in detail.

We tuned the models for each task separately, changing the maximum order of character and word n-gram features, case normalization, and SVM margin parameter 'C'. The parameter ranges explored during tuning was 0–12 for maximum character n-gram order, 0–7 for maximum word n-gram order, and 0.1–2.0 with steps of 0.1 for 'C'. We used 5-fold cross validation during tuning, using random search through the space of hy-

Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task, pages 52–53
Brussels, Belgium, October 31, 2018.

Task	tf-idf		BM25	
	devel.	test	devel.	test
1	90.17	90.87	90.13	91.22
2	76.42	46.8	76.45	46.5
3	93.52	40.4	93.42	42.4
4 (train)	89.22	–	89.41	–
4 (full)	90.16	85.53	90.16	85.53

Table 1: F1-scores of tf-idf and BM25 weighted models on the development set and the official test set. The F1-scores for task 2 are micro-averaged. The two set of scores for Task 4 reflect the difference between the full labeled-data set (including additional 1211 training instances) in comparison to the original training set.

perparameters described above. Approximately 1000 random hyperparameter settings were tried for each model. The models with the best parameter settings were retrained using the complete training data for producing the final predictions.

The source of the texts for all tasks is Twitter. At the time we downloaded them, some tweets were not available, resulting in training set sizes of 9182, 15 723, 16 888, and 5759 for tasks 1, 2, 3 and 4 respectively. Some of these numbers are substantially lower than that of intended number of training samples of 10 000, 17 000, 25 000, and 8180 respectively. For task 4, we also used an additional 1211 tweets, initially planned as the test set for this task. The test sets contained (approximately) 5000 tweets for tasks 1, 2 and 3, and a considerably smaller number (161) for task 4. All training sets showed some degree of class imbalance. The imbalance was particularly strong for tasks 3 and 4, where over 70 % and 90 % of the instances belonged to the negative class, respectively. Further information on the data sets can be found in Weissenbacher et al. (2018).

3 Results and Discussion

Table 1 presents F_1-scores of the models on each task. In general, we do not observe substantial differences between the term weighting schemes, but for some tasks the gap between training and development set scores is rather large. We do not know the system rankings at the time of writing, but only know that the results above are above the mean of the best-scores from all participating teams.

The systems we used for the shared task are simple, yet, effective classifiers with character and word n-gram features. The big discrepancies between the development and test set scores in task 2 and task 3 points either some differences between the distributions of the training and test sets, or it may also be due to large amount of missing tweets in our training set, indicating more data is likely to be particularly useful in these tasks. We also compared the effectiveness of two feature weighting systems, tf-idf and BM25, which did not show any substantial differences. Since our models were originally intended as baseline models, the scores presented in Table 1 were obtained without the use of any external data or source of information. Better results are likely by use of external information, such as appropriate dictionaries, term lists, or embeddings trained on large amounts of unlabeled data.

References

Çağrı Çöltekin and Taraka Rama. 2017. Tübingen system in vardial 2017 shared task: experiments with language identification and cross-lingual parsing. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 146–155, Valencia, Spain.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs at emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, second edition. Pearson Prentice Hall.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Taraka Rama and Çağrı Çöltekin. 2017. Fewer features perform well at native language identification task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 255–260, Copenhagen, Denmark.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at EMNLP 2018. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.