BIGODM System in the Social Media Mining for Health Applications Shared Task 2019

Chen-Kai Wang[1], Hong-Jie Dai, PhD[2], Bo-Hung Wang[2]
[1]Big Data Laboratories, Chunghwa Telecom Labopratories, Taoyuan, Taiwan, R.O.C.
denniscckwang@gmail.com
[2]Department of Electrical Engineering,
National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, R.O.C.
{hjdai, 1061247131}@nkust.edu.tw

Abstract

In this study, we describe our methods to automatically classify Twitter posts conveying events of adverse drug reaction (ADR). Based on our previous experience in tackling the ADR classification task, we empirically applied the vote-based under-sampling ensemble approach along with linear support vector machine (SVM) to develop our classifiers as part of our participation in ACL 2019 Social Media Mining for Health Applications (SMM4H) shared task 1. The best-performed model on the test sets were trained on a merged corpus consisting of the datasets released by SMM4H 2017 and 2019. By using VUE, the corpus was randomly under-sampled with 2:1 ratio between the negative and positive classes to create an ensemble using the linear kernel trained with features including bag-of-word, domain knowledge, negation and word embedding. The best performing model achieved an F-measure of 0.551 which is about 5% higher than the average F-scores of 16 teams.

1 Introduction

Our team participated in the Social Media Mining for Health Applications (SMM4H) shared Task 1, which focus on the task of automatic classification of adverse effects mentions in tweets to distinguish tweets mentioned adverse effect (AE) from others(Weissenbacher et al., 2019).

2 Methods

AEC (Adverse Effect Classification) task is a typical classification problem. We used support vector machine (SVM) with the linear kernel to develop our classifiers. The training and validation sets released by the SMM4H 2019 organizers include 24,861 and 5,000 tweets, respectively. The organizers provided the entire training set but the validation set was downloaded by ourselves using the Twitter API. Unfortunately, only 2,887 tweets in the validation set can be downloaded from the Twitter website. In addition, we included the corpus released in AMIA-SMM4H 2017 for the same purpose, which contains 11,564 tweets (SarkerandGonzalez-Hernandez, 2017). We merged the two datasets and filtered out duplicate tweets to create a merged corpus for our model. The merged corpus contains 3,423 positive tweets and 31,858 negative tweets.

The imbalance ratio for the compiled corpus is 9.3, which is highly imbalanced. In order to develop classifiers with reliable performance, we implemented a vote-based under-sampling ensemble (VUE) technique Wang et al. (2018). VUE exploits all training examples in majority (negative) cases with under-sampling for creating an ensemble of SVM classifiers. It samples several subsets from the negative tweets without replacement and then create an ensemble by using each subset along with the minority cases (positive). The prediction can be determined by taking a majority vote among the separately created classifiers.

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 117–119
Florence, Italy, August 2, 2019.

In order to extract features for training our classifiers, we first pre-processed tweets to replace URLs, dosages and Twitter specific characters with the corresponding symbols, and modified the numeral parts in each token to one as proposed in our previous work Dai et al. (2016). The preprocessed tweet was then processed by a tweet tokenizer (Owoputi et al., 2013) to generate tokens. Follow by the above step, each token was processed by Hunspell to detect spelling errors. If a token is considered to be misspelled, the first recommended correction is included as an alternative term for the token. Finally, we lowercased all tokens and used the Snowball stemmer(Porter, 2001) to perform stemming without removing any stop words.

After the above steps, we extracted the following features to train our SVM models:

- Bag-of-word features: we extracted unigram and bigram with TF-IDF (Term Frequency-Inverse Document Frequency) as the weighting scheme.
- Domain knowledge features: The presence of adverse drug reaction (ADR) or drug mentions were engineered as two binary features with the value of either 0 or 1. The occurrences of ADR and drug names were recognized by using the ADR mention recognizer developed in our previous work Dai et al. (2016) and Wang et al. (2018).
- Negation features: The feature set uses three flags to indicate the occurrence of an ADR mention is missing, positive or negated. If a tweet contains ADRs, the NegEx algorithm (Chapman et al., 2001) is employed to determine whether the occurrence is negated.
- Word embedding features: The word embedding features proposed in our previous work Wang et al. (2018) was developed. The features were generated by taking the mean across all tokens' embedding represented as a 400-dimensional vector based on the pre-trained tweet WE model released by (Godin et al., 2015).

3 Results

Figure 1 show the results of the 10-fold cross validation (CV) on the training set of the AEC task. The standard precision (P), recall (R) and F-measure (F) are used to report the performance. Configuration 1 is the VUE model trained with the developed features. After submitting the results, we developed configuration 2, which was a baseline model with the same features but didn't apply any imbalanced techniques. The above two configurations of the developed classifiers were trained on the following three corpora:

1. SMM4H 2017 corpus
2. SMM4H 2019 corpus
3. SMM4H 2017+SMM4H 2019 corpus

During the participating the AEC task, we used configuration 1 with the above three corpora to conduct ablation experiments and submitted three runs corresponding to the first three configurations shown in Figure 1. The experimental results show that the VUE method has better recall but lower precision. The F-scores of VUE are better than baseline on the first two corpora. It is interesting to see that the baseline configuration performs better than VUE on the merged corpus.

4 Conclusion

In this paper, we briefly describe our systems developed for the SMM4H 2019 AEC task. Our best submitted run was based on the VUE model trained on a merged corpus. However, we noticed that by using the merged corpus, the baseline model which didn't exploit imbalanced technique performs better than that of VUE on the 10 fold CV. We will conduct error analysis to investigate the interesting results and compare the performance of other advanced imbalance techniques developed in our previous work Dai and Wang (2019).

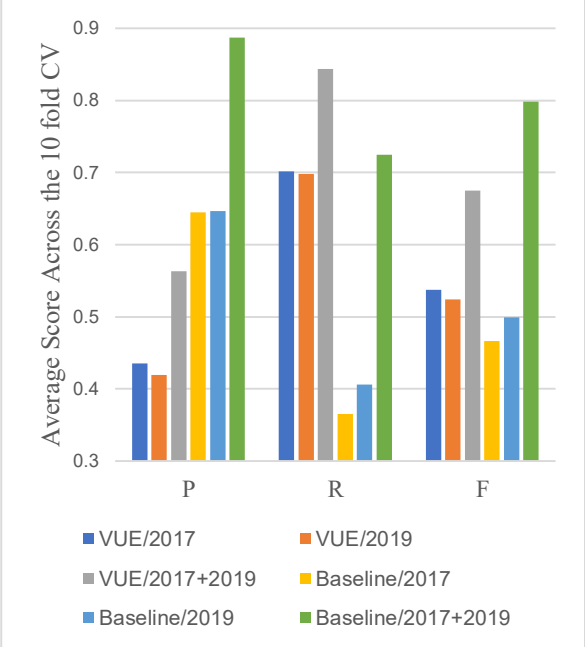

Figure 1: 10 fold CV on the training set of the AEC task.

References

Chapman, Wendy W, Bridewell, Will, Hanbury, Paul, Cooper, Gregory F, & Buchanan, Bruce G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics, 34*(5), 301-310.

Dai, Hong-Jie, Touray, Musa, Jonnagaddala, Jitendra, & Syed-Abdul, Shabbir. (2016). Feature engineering for recognizing adverse drug reactions from twitter posts. *Information, 7*(2), 27.

Dai, Hong-Jie, & Wang, Chen-Kai. (2019). Classifying Adverse Drug Reactions from Imbalanced Twitter Data. *International Journal of Medical Informatics*.

Godin, Fréderic, Vandersmissen, Baptist, De Neve, Wesley, & Van de Walle, Rik. (2015). *Multimedia Lab $@ $ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations.* Paper presented at the Proceedings of the Workshop on Noisy User-generated Text.

Owoputi, Olutobi, O'Connor, Brendan, Dyer, Chris, Gimpel, Kevin, Schneider, Nathan, & Smith, Noah A. (2013). *Improved part-of-speech tagging for online conversational text with word clusters.* Paper presented at the Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies.

Porter, Martin F. (2001). Snowball: A language for stemming algorithms. In.

Sarker, Abeed, & Gonzalez-Hernandez, Graciela. (2017). Overview of the Second Social Media Mining for Health (SMM4H) shared tasks at AMIA 2017. *Training, 1*(10,822), 1239.

Wang, Chen-Kai, Dai, Hong-Jie, Wang, Feng-Duo, & Su, Emily Chia-Yu. (2018). *Adverse Drug* Reaction *Post Classification with Imbalanced Classification Techniques.* Paper presented at the 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI).

Weissenbacher, Davy , Sarker, Abeed, Magge, Arjun, Daughton, Ashlynn, O'Connor, Karen, Paul, Michael, & Graciela, Gonzalez-Hernandez. (2019). *Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Task at ACL 2019.* Paper presented at the Proceedings of the 2019 ACL Workshop SMM4H: The 4th Social Media Mining for Health Applications Workshop & Shared Task.