ISLab System for SMM4H Shared Task 2020

Chen-Kai Wang[1,2], You-Chen Zhang[3], Bo-Chun Xu[3], Bo-Hong Wang[3], You-Ning Xu[3], Po-Hao Chen[3], Hong-Jie Dai[3,4,5*], Chung-Hong Lee[3]

[1] Big Data Laboratories, Chunghwa Telecom Laboratories, Taoyuan, Taiwan, R.O.C

[2] Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

[3] Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, R.O.C.

[4] Department of Post-Baccalaureate Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan, R.O.C.

[5] National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan, R.O.C.

{denniscкwang, magicpower503, xobochun0116}@gmail.com, {1061247131, 108111109, f107154153, hjdai}@nkust.edu.tw, leechung@mail.ee.kuas.edu.tw

Abstract

In this paper, we described our systems for the first and second subtasks of Social Media Mining for Health Applications (SMM4H) shared task in 2020. The two subtasks are automatic classification of medication mentions and adverse effect in tweets. Our systems for both subtasks are based on Robustly optimized BERT approach (RoBERTa) and our previous work at SMM4H'19. The best F1-scores achieved by our systems for subtask 1 and 2 were 0.7974 and 0.64 respectively, which outperformed the average F1-scores among all teams' best runs by at least 0.13.

1 Introduction

Nowadays, social media are often being used by general public to create and share public messages related to their health. With the global increase in social media usage, there is a trend of posting information related to adverse drug reactions (ADR). Mining social media data for this type of information is helpful for pharmacological post-marketing surveillance and monitoring. In order to facilitate the use of social media for health monitoring and surveillance, we participated in the social media mining for health applications (SMM4H) shared task to develop systems that can automatically identify tweets conveying medications and adverse effects.

2 Task and Data Description

2.1 Task 1: Automatic Classification of Tweets that Mention Medications

This task is a binary classification task involves distinguishing tweets that determine whether it mentions medications or dietary supplements. The organizers provided a training set consisting of 69,272 tweets for all participants to develop their system, and a test set consisting of 29,687 tweets. Table 1 shows the distribution of the binary labels over the training and test sets. We can find that the training set is highly imbalanced.

Table 1: Distributions of labels over the training, validation and test datasets of task 1.

Dataset	Positive (1)	Negative (0)	Total
Training set	146 (0.26%)	55,273 (99.74%)	55,419
Validation set	35 (0.25%)	13,818 (99.75%)	13,853
Test set	N/A	N/A	29,687

* Corresponding authors

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 42–45
Barcelona, Spain (Online), December 12, 2020.

2.2 Task 2: Automatic Classification of Multilingual Tweets that Report Adverse Effects

Task 2 is also a binary classification task, which involves distinguishing tweets mentions ADRs. As illustrated in Table 2, the training set of the task is also highly imbalanced.

Table 2: Distribution of labels over the training, validation and test datasets of task 2.

Dataset	Positive (1)	Negative (0)	Total
Training set	1,903 (9.26%)	18,641 (90.74%)	20,544
Validation set	474 (9.23%)	4,660 (90.77%)	5,134
Test set	N/A	N/A	4,759

3 Methods

For each task, we developed three systems; the first and the second was based on Robustly Optimized BERT Pretraining Approach (RoBERTa) (Y. Liu et al., 2019) and the third was based on the method proposed in our previous work Dai and Wang (2019); (Wang et al., 2019).

3.1 System 1 and 2: RoBERTa and Retrained RoBERTA

BERT is an unsupervised language representation method to obtain deep bidirectional representations of sentences by jointly conditioning on both left and right context in all layers from free text. RoBERTa (Y. Liu et al., 2019) is an enhanced version of BERT which was trained with dynamic masking, full sentence without next sentence prediction loss, large mini-batches and a larger byte-level byte-pair encoding (BPE) (Sennrich et al., 2015). BPE is a hybrid between character- and word-level representations allowing handling large vocabularies in natural language corpora. Radford et al. (2019) introduced a clever implementation of BPE by using bytes instead of Unicode characters as the base sub-word units, which makes it possible to learn a sub-word vocabulary of a modest size that can still encode any input text without introducing any unknown tokens. In our implementation, we encoded both datasets released by the SMM4H organizers through BPE, and fine-tuned the RoBERTa-large architecture pre-trained model on the released training set to develop our first system.

For the system 2, we retrained the RoBERTa-large architecture on a tweet unlabeled corpus collected by our team. The unlabeled corpus consisted of 6,339,457 tweets on Twitter collected from January to April 2019, according to 183,593 drug names recorded in RxNorm (S. Liu et al., 2005) and 13,699 ADRs released by Nikfarjam et al. (2015).

For both models, we set the same parameters to fine-tuned RoBERTa; each model was trained by using the Adam optimizer with a learning rate of 10^{-5} for 10 epochs with a batch size of 8.

3.2 System 3: Random Under Sampling with Word Embedding-based Synthetic Minority Over-sampling Technique

Because the datasets of both subtasks are highly imbalanced, we applied the word embedding-based synthetic minority over-sampling technique (WESMOTE) with Random Under Sampling (RUS) proposed in our previous work (DaiandWang, 2019; Wang et al., 2019) to develop classifiers with reliable performance. In our implementation, we first applied WESMOTE to synthesize new positive examples by using the sentence representation based on BERT. We then randomly under-sampled the negative examples so that the ratio of positive against negative is 1:2. In order to extract features for training our classifiers, we 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 (Anonymous, 2019) 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 support vector machine (SVM) model:

- 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; 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 features were generated by taking the mean across all tokens' embedding represented by a 1024-dimensional vector based on the whole word masking variant of BERT-Large released by Turc et al. (2019).

4 Results

Table 3 and Table 4 show the performance of our systems for task 1 and task 2 on the validation and test data, respectively. The pre-trained RoBERTa model achieved the best F1-scores on both tasks. Compared with RUS_WESMOTE and Retrained RoBERTa on the two validation sets, RoBERTa exhibited much better performance too.

Table 3: Performance on validation and test data for task 1.

System	Validation			Test		
	P	R	F	P	R	F
RoBERTa	**0.938**	**0.857**	**0.896**	0.803	0.792	**0.797**
RUS_WESMOTE	0.484	0.534	0.508			0.47
Retrained RoBERTA	0.558	0.686	0.615			0.48
Average scores				0.7032	0.6948	0.6628

Table 4: Performance on validation and test data for task 2.

System	Validation			Test		
	P	R	F	P	R	F
RoBERTa	**0.662**	**0.711**	**0.686**	0.62	0.65	**0.64**
RUS_WESMOTE	0.544	0.527	0.535			0.41
Retrained RoBERTA	0.481	0.627	0.544			0.45
Average scores				0.42	0.59	0.46

5 Discussion

For both subtasks, the distribution of binary class is highly imbalanced. When we used the traditional machine learning methods like SVM, even we have tried to deal with the data imbalance problem by RUS_WESMOTE, the performance is still far below that of the system with the pre-trained RoBERTa. On the other hand, for the RoBERTa based systems, we didn't apply any imbalance techniques but they still get compatible and even better precision, recall and F-score. It is surprising to see that the system based on the tweet-retrained RoBERTa model didn't outperform the original pre-trained RoBERTa model.

6 Conclusion

We demonstrated that the system based on the pre-trained RoBERTa model outperformed traditional SVM-based method and the re-trained RoBERTa model. We will conduct error analysis to interpret the results of the two RoBERTa-based models to figure out the reason why the performance of the re-trained RoBERTa model get worse in the future.

Acknowledgements

This study was supported by the Ministry of Science and Technology of Taiwan [Grant numbers MOST-106-2221-E-143-007-MY3] and [Grant numbers MOST 109-2221-E-992-074-MY3].

Reference

Anonymous. (2019). Hunspell. Retrieved from http://hunspell.github.io/

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, 129, 122-132.

Liu, Simon, Ma, Wei, Moore, Robin, Ganesan, Vikraman, & Nelson, Stuart %J IT professional. (2005). RxNorm: prescription for electronic drug information exchange. 7(5), 17-23.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, . . . Stoyanov, Veselin. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Nikfarjam, Azadeh, Sarker, Abeed, O'connor, Karen, Ginn, Rachel, & Gonzalez, Graciela %J Journal of the American Medical Informatics Association. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. 22(3), 671-681.

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, 2001. In.

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, & Sutskever, Ilya %J OpenAI Blog. (2019). Language models are unsupervised multitask learners. 1(8), 9.

Sennrich, Rico, Haddow, Barry, & Birch, Alexandra. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

Turc, Iulia, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. (2019). Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962.

Wang, Chen-Kai, Dai, Hong-Jie, & Wang, Bo-Hung. (2019). BIGODM System in the Social Media Mining for Health Applications Shared Task 2019. Paper presented at the Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task.

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).