Sentence Transformers and Bayesian Optimization for Adverse Drug Effect Detection from Twitter

Oguzhan Gencoglu
Tampere University, Faculty of Medicine and Health Technology, Tampere, Finland
oguzhan.gencoglu@tuni.fi

Abstract

This paper describes our approach for detecting adverse drug effect mentions on Twitter as part of the Social Media Mining for Health Applications (SMM4H) 2020, Shared Task 2. Our approach utilizes multilingual sentence embeddings (sentence-BERT) for representing tweets and Bayesian hyperparameter optimization of sample weighting parameter for counterbalancing high class imbalance.

1 Introduction and Related Work

Automatic adverse drug reaction detection from social media has high significance to health informatics and pharmacovigilance due to fast, scalable, and diverse public health surveillance opportunities. Numerous studies have proposed natural language processing and machine learning solutions to detect mentions of adverse drug effects, especially from Twitter (Bian et al., 2012; Jiang and Zheng, 2013; Ginn et al., 2014; O'Connor et al., 2014; Katragadda et al., 2015; Egger et al., 2016; Korkontzelos et al., 2016; Rastegar-Mojarad et al., 2016; MacKinlay et al., 2017; Alimova and Tutubalina, 2017; Moh et al., 2017; Lee et al., 2017; Gupta et al., 2018a; Gupta et al., 2018b; Masino et al., 2018; Wu et al., 2019; Mesbah et al., 2019; Zhang et al., 2019; Alhuzali and Ananiadou, 2019). Proposed approaches in earlier studies vary from rule-based systems to deep learning. The challenge of detecting adverse drug effect mentions in Twitter remains unsolved due to lack of large-scale annotated datasets, rareness of relevant tweets among all tweets as well as ever-changing nature of the phenomenon.

The task consists of building separate adverse drug effect mention detection (binary classification) models for English, French, and Russian datasets from Twitter (9.25%, 1.61%, and 8.75% positive class prevalence, respectively). Detailed description of data and task can be found in (Klein et al., 2020).

2 Methods

2.1 Sentence Embeddings

We first preprocess the tweets by removing the usernames (in the form `@username`) and urls. We utilize recently introduced sentence-BERT (SBERT) models (Reimers and Gurevych, 2019) to represent each tweet as a *sentence embedding* instead of using standard BERT (Devlin et al., 2019) or its variants which work in a token-embedding manner. SBERT models are trained by a Siamese-network structure using BERT models to learn semantically meaningful sentence embeddings in an efficient manner. We utilize sentence-embedding versions of pretrained RoBERTa (Liu et al., 2019) model for English dataset (representation vectors of length 1024) and multilingual DistilBERT (Sanh et al., 2019) model (trained on 13 languages) for French and Russian datasets (representation vectors of length 512). We then train 3-layer (2 dense layers of size 256 and 32 with *ReLU* activation, respectively and an output layer with *sigmoid* activation) fully-connected neural networks using the sentence embeddings as input features. A Dropout rate of 0.5 is used between the dense layers. Model trainings are performed in a mini-batch manner for 100 epochs with a batch size of 32 with *Adam* optimizer (learning rate of 5×10^{-4}). Every

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 161–164
Barcelona, Spain (Online), December 12, 2020.

run is performed in a 5-fold cross-validation manner and models at the epoch that maximizes the F_1 score on the validation split of the cross-validation is selected as final model.

2.2 Bayesian Optimization of Sample Weighting

As the problem at hand consists of highly imbalanced datasets, statistical balancing of positive and negative classes benefits the model training. Such balancing may be performed in various ways, e.g., with data augmentation. We chose to tackle the problem by increasing the contribution of the positive samples (corresponding to adverse effect mentions) to the loss function calculation. Each task requires a different weight coefficient (multiplier) for the positive samples, θ, as the prevalence of positive samples differ between datasets. We formulate the problem of finding the optimal multiplier, $\hat{\theta}$, as a Bayesian optimization problem:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(\theta), \tag{1}$$

where $f(\theta)$ is the average of cross-validation F_1 scores, i.e., $\frac{1}{N}\sum_{i=1}^{N} F_1^i$. For our experiments $N = 5$ as we perform 5-fold cross-validation. We use a *Gaussian Process* for the surrogate model (Rasmussen, 2003) of the Bayesian optimization by which we emulate the statistical relationship between the positive sample weight coefficient and cross-validation performance, given a dataset.

3 Results

Results of our experiments and submissions can be examined from Table 1. Validation results correspond to local experiments, i.e., cross-validation with training data and inference on validation data. Test results (competition results) correspond to cross-validation with training + validation data and inference on test data from the ensembles of the cross-validation models (*mean* pooling). F_1 scores of 0.48, 0.17, and 0.42 are achieved for English, French, and Russian datasets, respectively.

			Validation			**Test**			**Comp. Avg.**
Task	**SBERT model**	$\hat{\theta}_{best}$	**P**	**R**	$\boldsymbol{F_1}$	**P**	**R**	$\boldsymbol{F_1}$	$\boldsymbol{F_1}$
English	RoBERTa	2.77	0.46	0.54	0.49	0.44	0.53	0.48	0.46
French	DistilBERT	19.67	0.19	0.25	0.22	0.15	0.20	0.17	0.07
Russian	DistilBERT	8.23	0.38	0.52	0.45	0.35	0.55	0.42	0.43

Table 1: Validation and competition test results of developed binary classification models for the 3 datasets (P = Precision, R = Recall).

4 Discussion and Conclusions

We chose to use pre-computed sentence embeddings instead of fine-tuning on token embeddings out of original BERT models due to its low computational requirements (Reimers and Gurevych, 2019). While it is possible to represent tweets with a standard BERT model as well (e.g. by averaging the token embeddings out of the output layer), such sentence representations were shown to be inferior to embeddings specifically trained for representing sentences (Reimers and Gurevych, 2019).

Bayesian optimization is beneficial in settings where the function to be minimized/maximized is a black-box function without a known closed-form, expensive to evaluate, and stochastic (Močkus, 1975). As $f(\theta)$ corresponds to cross-validation performance in our case, it is indeed a black-box function and computationally expensive to evaluate. Furthermore, it is a stochastic function as the same θ may give different outputs for different runs due to randomness in neural network weight initialization. That is our motive for employing Bayesian hyperparameter optimization for sample weighting. Advantage of using a Gaussian Process as the surrogate model is that it can provide uncertainty estimations. As expected, optimum positive sample weight parameters found by Bayesian optimization are different for each dataset and French dataset requires much higher balancing due to its extreme class inbalance. Our Bayesian optimization approach can be easily extended to optimization of other hyperparameters of the model including neural network architecture and other relevant design choices.

References

Hassan Alhuzali and Sophia Ananiadou. 2019. Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 339–347.

Ilseyar Alimova and Elena Tutubalina. 2017. Automated detection of adverse drug reactions from social media posts with machine learning. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 3–15. Springer.

Jiang Bian, Umit Topaloglu, and Fan Yu. 2012. Towards large-scale Twitter mining for drug-related adverse events. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, pages 25–32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dominic Egger, Fatih Uzdilli, Mark Cieliebak, and L Derczynski. 2016. Adverse drug reaction detection using an adapted sentiment classifier. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. 2014. Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, pages 1–8. Citeseer.

Shashank Gupta, Manish Gupta, Vasudeva Varma, Sachin Pawar, Nitin Ramrakhiyani, and Girish Keshav Palshikar. 2018a. Co-training for extraction of adverse drug reaction mentions from tweets. In *European Conference on Information Retrieval*, pages 556–562. Springer.

Shashank Gupta, Sachin Pawar, Nitin Ramrakhiyani, Girish Keshav Palshikar, and Vasudeva Varma. 2018b. Semi-supervised recurrent neural network for adverse drug reaction mention extraction. *BMC Bioinformatics*, 19(8):212.

Keyuan Jiang and Yujing Zheng. 2013. Mining Twitter data for potential drug effects. In *International Conference on Advanced Data Mining and Applications*, pages 434–443. Springer.

Satya Katragadda, Harika Karnati, Murali Pusala, Vijay Raghavan, and Ryan Benton. 2015. Detecting adverse drug effects using link classification on Twitter data. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 675–679. IEEE.

Ari Z Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*.

Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H Gonzalez. 2016. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62:148–158.

Kathy Lee, Ashequl Qadir, Sadid A Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. 2017. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 705–714.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. version 1.

Andrew MacKinlay, Hafsah Aamer, and Antonio Jimeno Yepes. 2017. Detection of adverse drug reactions using medical named entities on Twitter. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1215. American Medical Informatics Association.

Aaron J Masino, Daniel Forsyth, and Alexander G Fiks. 2018. Detecting adverse drug reactions on Twitter with convolutional neural networks and word embedding features. *Journal of Healthcare Informatics Research*, 2(1-2):25–43.

Sepideh Mesbah, Jie Yang, Robert-Jan Sips, Manuel Valle Torre, Christoph Lofi, Alessandro Bozzon, and Geert-Jan Houben. 2019. Training data augmentation for detecting adverse drug reactions in user-generated content. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2349–2359.

Jonas Močkus. 1975. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer.

Melody Moh, Teng-Sheng Moh, Yang Peng, and Liang Wu. 2017. On adverse drug event extractions using Twitter sentiment analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 6(1):18.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on Twitter? mining tweets for adverse drug reactions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 924. American Medical Informatics Association.

Carl Edward Rasmussen. 2003. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Yue Yu, and Hongfang Liu. 2016. Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. version 4.

Chuhan Wu, Fangzhao Wu, Zhigang Yuan, Junxin Liu, Yongfeng Huang, and Xing Xie. 2019. MSA: Jointly detecting drug name and adverse drug reaction mentioning tweets with multi-head self-attention. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 33–41.

Tongxuan Zhang, Hongfei Lin, Yuqi Ren, Liang Yang, Bo Xu, Zhihao Yang, Jian Wang, and Yijia Zhang. 2019. Adverse drug reaction detection via a multihop self-attention mechanism. *BMC Bioinformatics*, 20(1):479.