NCUEE-NLP@SMM4H'22: Classification of Self-reported Chronic Stress on Twitter Using Ensemble Pre-trained Transformer Models

Tzu-Mi Lin, Chao-Yi Chen, Yu-Wen Tzeng and Lung-Hao Lee
Department of Electrical Engineering, National Central University, Taiwan

Abstract

This study describes our proposed system design for the SMM4H 2022 Task 8. We fine-tune the BERT, RoBERTa, ALBERT, XLNet and ELECTRA transformers and their connecting classifiers. Each transformer model is regarded as a standalone method to detect tweets that self-reported chronic stress. The final output classification result is then combined using the majority voting ensemble mechanism. Experimental results indicate that our approach achieved a best F1-score of 0.73 over the positive class.

1 Introduction

The Social Media Mining for Health Application (SMM4H) shared tasks involve NLP challenges on social media data for health research. We participated in the SMM4H 2022 Task 8 (Weissenbacher et al., 2022), focusing on automatically differentiating tweets that exhibit self-reported chronic stress (annotated as "1") from those that do not (annotated as "0"). Chronic stress is a kind of long-term physiological or psychological response which may damage mental health. Deep sparse neural networks were used to detect stress on social media data (Lin et al., 2014). Factor graph models were combined with convolutional neural networks to detect user stress (Lin et al., 2017). An emotion-infused language model was proposed to facilitate stress detection (Turcan et al., 2021). Emotion recognition was investigated as an auxiliary task to improve stress detection (Yao et al., 2021). Recently, novel transformer-based neural networks have achieved highly promising results in many NLP tasks. This trend motivates us to explore the use of pre-trained transformer models to detect chronic stress.

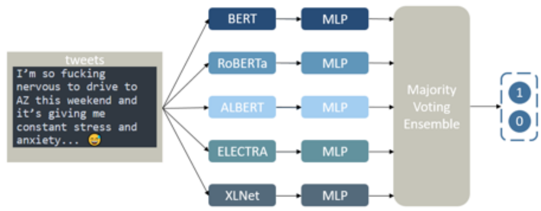

Figure 1: NCUEE-NLP system architecture.

This paper describes a system proposed by the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing Lab) for the SMM4H 2022 Task 8 using transformer models including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), XLNet (Yang et al., 2019), and ELECTRA (Clark et al., 2020), followed by fine-tuning for the classification task. In addition, a majority voting ensemble mechanism is used to enhance system performance. The evaluation metric is F1-score for the positive class (i.e., tweets annotated as "1"). Our best F1-score is 0.73 over the positive class.

2 The NCUEE-NLP System

Figure 1 shows our NCUEE-NLP system architecture for the SMM4H 2022 shared task 8. Our system is composed of two parts: 1) pre-trained transformer models; and 2) ensemble mechanism.

We approach the task using five pre-trained transformer models, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), XLNet (Yang et al., 2019), and ELECTRA (Clark et al., 2020). We use training, validation and test datasets to fine-tune the

Proceedings of the 29th International Conference on Computational Linguistic, pages 62–64
October 12–17, 2022.

Pre-trained Transformer Models		Validation set		
		precision	recall	F1-score
Standalone	BERT (B)	0.6567	0.8461	0.7395
	RoBERTa (RB)	0.6823	0.8397	0.7529
	ALBERT (AB)	0.7172	0.6667	0.6910
	XLNet (XL)	0.6683	0.8654	0.7542
	ELECTRA (ET)	0.7076	0.7756	0.7400
Ensemble	B + RB + AB + XL + ET	0.6984	0.8462	0.7652
	B + RB + AB	0.6919	0.8205	0.7507
	B + RB + XL	0.6750	0.8654	0.7584
	B + RB + ET	0.7027	0.8333	0.7625
	B + AB + XL	0.6915	0.8333	0.7558
	B + AB + ET	**0.7151**	0.7885	0.7500
	B + XL + ET	0.6939	**0.8718**	**0.7727**
	RB + AB + XL	0.6825	0.8269	0.7478
	RB + AB + ET	0.7118	0.7756	0.7423
	RB + XL + ET	0.6963	0.8526	0.7666
	AB + XL + ET	0.7022	0.8013	0.7485

Table 1: Submission results on the SMM4H 2022 Task 8 validation dataset.

language model of each transformer thereby improving the embedding representation. Individual language models are then connected to the Multi-Layer Perceptron as a classifier.

The ensemble mechanism uses multiple learning models to obtain better classification performance. We use majority voting ensemble (Lee et al., 2021), in which each transformer model makes an independent classification (i.e., a vote 0 or 1) for each testing instance. The final system prediction output is the one that receives a majority of votes.

3 Evaluation

The experimental datasets were mainly provided by task organizers (Yang et al., 2022), with a total of 2,936 tweets in the training set, including 1,092 positive and 1,844 negative tweets. The validation set contains 534 tweets (156 positive/264 negative).

All tweets were pre-processed using the following steps: 1) Converting username into @USER; 2) Converting emojis into textual representation; 3) Removing all URLs; 4) Removing '#' from hashtags; 5) Segmenting hashtags which contained two or more words into their component words; and 6) Removing multi-lingual characters to retain English letters and numbers only.

The pre-trained transformer models were downloaded from HuggingFace (Wolf et al., 2019). We selected an odd number (i.e., three of five) of models for majority voting ensemble. A total of 12 groups was combined for the ensemble mechanism. The hyper-parameters used were as follows: training batch size 100, learning rate 1e-5, and maximum sequence length 256.

Table 1 shows the results on the SMM4H 2022 Task 8 validation set. Among individual transformer models, ALBERT clearly underperformed the other four models which had similar F1-score results. The five ensemble transformer models enhanced performance compared with the best standalone model, RoBERTa. Among all three ensemble models, those excluding ALBERT resulted in some performance improvements, with the best combination of BERT, XLNet and ELECTRA, achieving an F1-score of 0.7727.

We selected the top three ensemble models that performed well on the validation set as our final testing models. The combination of RoBERTa, XLNET and ELECTRA performed achieved the best F1-score of 0.73 on the test set.

4 Conclusions

This study describes the NCUEE-NLP system submission in SMM4H 2022 Task 8 for self-reported chronic stress, including system design, implementation and evaluation. Our submitted ensemble pre-trained transformer models achieved a high F1-score of 0.73.

Acknowledgments

This study was partially supported by the National Science and Technology Council, Taiwan under the grant MOST 111-2628-E-008-005-MY3.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the 8th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2003.10555

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. https://doi.org/10.48550/arXiv.1810.04805

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.1909.11942

Lung-Hao Lee, Yuh-Shyang Wang, Chao-Yi Chen, and Liang-Chih Yu. 2021. Ensemble Multi-Channel Neural Networks for Scientific Language Editing Evaluation. *IEEE Access,* 9:158540 – 158547. https://doi.org/10.1109/ACCESS.2021.3130042.

Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Jie Huang, Lianhong Cai, and Ling Feng. 2014. Psychological stress detection from cross-media microblog data using Deep Sparse Neural Network. In *Proceedings of the 2014 IEEE International Conference on Multimedia and Expo*, pages 1-6. https://doi.org/10.1109/ICME.2014.6890213

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017. Detecting Stress Based on Social Interactions in Social Networks. *IEEE Transactions on Knowledge and Data Engineering,* 29(9):1820-1833. https://doi.org/10.1109/tkde.2017.2686382

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*, https://arxiv.org/abs/1907.11692

Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. Emotion-Infused Models for Explainable Psychological Stress Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909. https://doi.org/10.18653/v1/2021.naacl-main.230

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, Graciela Gonzalez-Hernandez. 2022. Overview of the Seventh Social Media Mining for Health Applications #SMM4H Shared Tasks at COLING 2022. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's transformers: state: of-the-art natural language processing. *arXiv preprint*. https://doi.org/10.48550/arXiv.1910.03771

Yuan-Chi Yang, Angle Xie, Sangmi Kim, Jessica Hair, Mohammed Ali Al-Garadi, and Abeed Sarker. 2022. Automatic detection of twitter users who express chronic stress experiences via supervised machine learning and natural language processing. *Computers, Informatics, Nursing*. article in press.

Yiqun Yao, Michalis Papakostas, Mihai Burzo, Mohamed Abouelenien, and Rada Mihalcea. 2021. MUSER: MUltimodal Stress detection using Emotion Recognition as an Auxiliary Task. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2714–2725. https://doi.org/10.18653/v1/2021.naacl-main.216

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1906.08237