Continuing Pre-trained Model with Multiple Training Strategies for Emotional Classification

Bin Li[1]*, Yixuan Weng[2]*, Qiya Song[1]*, Bin Sun[1], Shutao Li[1]
[1] College of Electrical and Information Engineering, Hunan University
[2] National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy Sciences, Beijing, 100190, China
{libincn, sqyunb, sunbin611, shutao_li}@hnu.edu.cn, wengsyx@gmail.com

Abstract

Emotion is the essential attribute of human beings. Perceiving and understanding emotions in a human-like manner is the most central part of developing emotional intelligence. This paper describes the contribution of the LingJing team's method to the Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) 2022 shared task on Emotion Classification. The participants are required to predict seven emotions from empathic responses to news or stories that caused harm to individuals, groups, or others. This paper describes the continual pre-training method for the masked language model (MLM) to enhance the DeBERTa pre-trained language model. Several training strategies are designed to further improve the final downstream performance including the data augmentation with the supervised transfer, child-tuning training, and the late fusion method. Extensive experiments on the emotional classification dataset show that the proposed method outperforms other state-of-the-art methods, demonstrating our method's effectiveness. Moreover, our submission ranked Top-1 with all metrics in the evaluation phase for the Emotion Classification task.

1 Introduction

Emotion is an important component of human daily communication. However, with the growing interest in human-computer interfaces, machines still lag in possessing and perceiving emotions. Understanding human emotional states in dialogue is crucial for building natural human-machine interaction, which aims to generate appropriate responses.

Emotion classification (EMO) in the text is concentrated on projecting words, sentences, and documents to a set of emotions according to psychological models proposed by (Ekman, 1992), which is an interdisciplinary field of study that span psychology and computer science. This task has evolved from a purely research-oriented topic to play a role in various applications, including mental health assessment, intelligent agents, social media mining (Calvo et al., 2017; Rambocas and Pacheco, 2018). Therefore, emotion classification has become a hot topic in the field of natural language processing (NLP), and lots of research efforts have been devoted to its development.

With the rapid development of artificial intelligence technology, especially deep learning, researchers have made substantial progress on EMO tasks over the past few decades. Before the era of deep learning, traditional EMO methods not only ignore the order of occurrence of words in written text, but are also limited by fixed input sizes. However, obtaining contextual relations between words from the sequence texts plays a crucial role in understanding the complete meaning of sentences. With the popularity of data-driven techniques, deep learning based methods improve the shortcomings of traditional methods and achieve superior EMO performance (Ran et al., 2018; Rajabi et al., 2020; Nandwani and Verma, 2021).

More recently, the transformer self-attention architecture based (Vaswani et al., 2017) pre-trained models have been successfully applied for learning language representations by exploiting large amounts of unlabeled data. These models mainly include BERT (Devlin et al., 2018a), OpenAI GPT (Radford et al., 2018), RoBERTa (Liu et al., 2019). These architectures show superior performance when fine-tuning different downstream tasks, including machine translation (Imamura and Sumita, 2019), text classification (Sun et al., 2019), emotion classification (Luo and Wang, 2019) and question answering (Garg et al., 2020). Recent works have shown that transformer-based pre-trained methods can achieve state-of-the-art performance in EMO tasks (Acheampong et al., 2021; Luo and Wang, 2019). Motivated by this, we adopt the DeBERTa model (He et al., 2020) with continual pre-training

*These authors contribute equally to this work.

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 233 - 238

method for the masked language model (MLM) (Devlin et al., 2018b) in this Track 2 to improve final downstream performance. More feasible training strategies are designed to improve the final results further. In this paper, we describe our work for Track 2 of the WASSA Shared Task 2022, addressing the issue of emotion classification.

2 Method

In this section, we will elaborate on the main methods for Track 2 of the WASSA 2022 Shared Task. More details about the training strategies are detailed at the end of this section.

2.1 Continuing Pre-training

It is a wise choice for further continual pre-training (Gururangan et al., 2020) to enhance the pre-trained model, i.e., DeBERTa model (He et al., 2020). It will be helpful to alleviate the task and domain discrepancy between the upstream and the downstream tasks (Qiu et al., 2020). As a result, we adopt the continual pre-training method for the masked language model (MLM) (Devlin et al., 2018b) in this Track 2 to directly improve final downstream performance. The available datasets are chosen from the open-source resources (Demszky et al., 2020; Öhman et al., 2020). The optimization function is written as follows

$$\max_{\theta} \log p_{\theta}(\boldsymbol{X} \mid \tilde{\boldsymbol{X}}) = \max_{\theta} \sum_{i \in \mathcal{C}} \log p_{\theta}\left(\tilde{x}_i = x_i \mid \tilde{\boldsymbol{X}}\right) \tag{1}$$

where the $\mathcal{C}$ is the index set of the masked tokens in the sequence.

We adopt the implementation of the original paper (Devlin et al., 2018b) to keep 10% of the masked tokens unchanged, another 10% replaced with randomly picked tokens and the rest replaced with the [MASK] token.

2.2 Emotion Classification with DeBERTa Model

Track 2 is a classic emotion classification task, where seven emotional labels are required to be classified. We adopt the DeBERTa-v2 (He et al., 2020) model with continuing pre-training method for processing this classification task, where the main method structure is shown in Figure 1. The given sentence is separated into tokens and then sent to the pre-trained language model (PLM) as the input. To obtain the complete meaning of the whole sentence, we take the output embedding of each token to be averaged by the averaged pooling layer. The seven-categories task is designed by passing the averaged encoding into the fully connected layer with dropout.

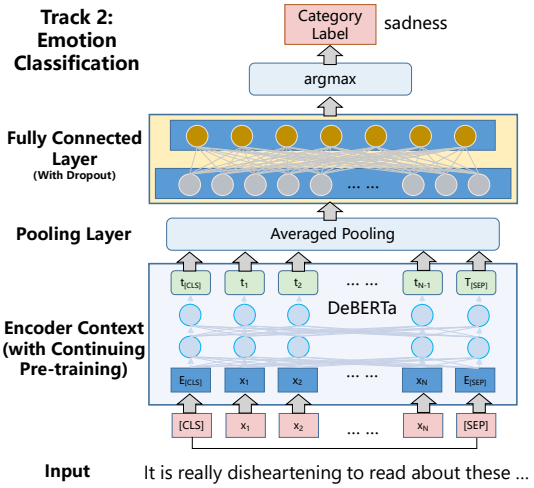

Figure 1: Main structure of the method in Track 2.

2.3 Training Strategies

We introduce some training strategies used in the Track 2 emotional classification, where the data augmentation with supervised transfer, child-tuning training, and late fusion will be introduced in detail.

2.3.1 Data Augmentation with Supervised Transfer

When fine-tuning on the English emotional classification datasets, we shall transfer the supervised knowledge into the Track 2 emotional task from the other datasets. Specifically, inspired by the work (Kulkarni et al., 2021), we adopt the data augmentation strategies with Random Augmentation (RA) and Balanced Augmentation (BA), where the GoEmotions (Demszky et al., 2020) and the XED dataset (Öhman et al., 2020) are adopted for implementation. It provides more useful knowledge transferred from the same resources to the downstream task (Durrani et al., 2021). As a result, the continuing pre-trained DeBERTa model fine-tuned on these similar datasets in English may achieve better results.

2.3.2 Child-tuning Training

The efficient Child-tuning (Xu et al., 2021) method is used for fine-tuning the DeBEATa model, where the parameters of the Child network are updated with the gradients mask. For the Track 2 task, the task-independent algorithm is used. In the phase of the fine-tuning, the gradient masks are obtained by

Bernoulli distribution (Chen and Liu, 1997) sampling from in each step of iterative update, which is equivalent to randomly dividing a part of the network parameters when updating. The equation of the above steps is shown as follows

$$w_{t+1} = w_t - \eta \frac{\partial \mathcal{L}(w_t)}{\partial w_t} \odot M_t$$
$$M_t \sim \text{Bernoulli}\ (p_F) \quad (2)$$

where the notation $\odot$ represents the dot production, p_F is the partial network parameter.

2.4 Late Fusion

Due to the complementary performance between different emotion prediction models (Colnerič and Demšar, 2018), we design the late fusion method with the Bagging algorithm (Breiman, 1996) to vote on the results of the various models. The Bagging algorithm is used during the prediction, which can effectively reduce the variance of the final prediction by bridging the prediction bias of different models, augmenting the overall generalization ability of the system.

3 Experimental Setting

This section will subsequently present emotion dataset, our experimental models, experimental settings, control of variables experiment.

3.1 Dataset

Computational detection and understanding of empathy is an important factor in advancing human-computer interaction (Liu, 2015). Buechel et al. (2018) presented the first publicly available gold standard for the text-based empathy prediction[1]. Two researchers collected articles from news websites. After that, they asked the participants to read the article. Moreover, participants were asked to rate their level of urgency and distress before describing their ideas and feelings about it in writing.

Each participant rating 6 items for empathy (e.g., warm,tender, moved) and 8 items for distress (e.g., trou-bled, disturbed, alarmed) using a 7-point scale for each of those. The final data set has 1860 samples in total. The author obtains their gold scores by averaging the submissions from different participants.

[1]Data and code are available at: https://github.com/wwbp/empathic_reactions

3.2 Implementation Details

We train the model using the Pytorch[2] (Paszke et al., 2019) on the NVIDIA A100 GPU and use the hugging-face[3] (Wolf et al., 2020) framework. For all uninitialized layers, We set the dimension of all the hidden layers in the model as 1024. The AdamW(Loshchilov and Hutter, 2018) optimizer which is a fixed version of Adam (Kingma and Ba, 2014) with weight decay, and set β_1 to 0.9, β_2 to 0.99 for the optimizer. We set the learning rate to 1e-6 with the warm-up (He et al., 2016). The batch size is 1. We set the maximum length of 512, and delete the excess. Linear decay of learning rate and gradient clipping is set to 1e-6. Dropout (Srivastava et al., 2014) of 0.1 is applied to prevent over-fitting. All experiments select the best parameters in the valid set. Finally, we report the score of the best model (valid set) in the test set.

We use the DeBERTa-v2-xxl (He et al., 2021) as our pre-trained model, and fine-tune the model. The DeBERTa[4] model comes with 48 layers and a hidden size of 1536. The total parameters are 1.5B, and it is trained with 160GB raw data. We spent three weeks on this continuing pre-training step.

3.3 Comparison with Baseline Methods

We compare our methods with Baseline methods on the datasets (Buechel et al., 2018). Results of comparative methods are reported on website[5]. IITK@WASSA (Mundra et al., 2021) fine-tuned the ELECTRA model with ensemble method. The [CLS] token was passed through a single linear layer to produce a vector of size 7, representing class probabilities. Moreover, they save the snapshots with the best validation scores.

Phoenix's approach(Butala et al., 2021) is primarily based on T5 Model (Raffel et al., 2020) or conditional generation of emotion labels. Hence before feeding into the network, the emotion prediction task is cast as feeding the essay text as input and training it to generate target emotion labels as text. This allows for the use of the same model, loss function, and hyper-parameters for the task of emotion prediction as is done in other Text Generation tasks.

[2]https://pytorch.org

[3]https://github.com/huggingface/transformers

[4]microsoft/deberta-v2-xxlarge

[5]https://competitions.codalab.org/competitions/28713#results

Methods	Macro F1	Micro F1	Accuracy	Macro Precison	Macro Recall	Micro Precision	Micro Recall
MTL (Fornaciari et al., 2021)	0.483	0.585	0.585	0.546	0.47	0.585	0.585
T5 (Butala et al., 2021)	0.502	0.594	0.594	0.550	0.483	0.594	0.594
ELECTRA-Ensemble (Mundra et al., 2021)	0.588	0.655	0.655	0.603	0.584	0.655	0.655
Ours	**0.698**	**0.754**	**0.754**	**0.740**	**0.679**	**0.754**	**0.754**

Table 1: Comparison with state-of-the-art methods. The best results are in bold.

Team Name	Macro F1	Micro F1	Accuracy	Macro Precison	Macro Recall	Micro Precision	Micro Recall
mantis	0.548	0.632	0.632	0.594	0.528	0.632	0.632
SURREY-CTS-NLP	0.548	0.634	0.634	0.576	0.532	0.634	0.634
SINAI	0.553	0.636	0.636	0.589	0.535	0.636	0.636
IUCL	0.572	0.646	0.646	0.599	0.555	0.646	0.646
Ours	**0.698**	**0.754**	**0.754**	**0.740**	**0.679**	**0.754**	**0.754**

Table 2: Results of the Top-5 teams participating in the EMO track for the post-evaluation. The best results are in bold.

Methods	Macro F1	Micro F1	Accuracy
Ours	**0.698**	**0.754**	**0.754**
w/o continuing pre-training	0.639	0.678	0.678
w/o supervised transfer	0.652	0.696	0.696
w/o child-tuning	0.656	0.699	0.699
w/o late fusion	0.664	0.664	0.664
w/o PLM	0.254	0.312	0.312

Table 3: Results of the ablation study.

Given the availability of further dependent variables (Fornaciari et al., 2021), create a Multi-Task Learning (MTL) model that takes the text as only input and jointly predicts emotions (classification task with categorical cross-entropy), empathy, and distress (regression task) (MTL2). They implemented a MIMTL model with text, gender, income, and IRI as input to predict emotions, empathy, and distress (MI3-MTL2).

4 Results and Discussions

The experiment results of the various methods on the evaluation dataset are displayed in Table 1. As presented in Table 1, our method achieves the best results in all evaluation metrics. Compared with the method from team IITK@WASSA that was the Top-1 last year, the adopted method gets a 0.110 increase of Macro F1, 0.137 increase of Macro Precision, 0.095 increase of Macro Recall, and 0.099 increase of Micro F1, Micro Precision, Micro Precision, and Accuracy. From this, we conclude that the proposed method outperforms the previous state-of-the-art method by an appreciable margin. It demonstrates the effectiveness of our method.

The Results of Top-5 teams participating in the EMO track for the post-evaluation are shown in Table 2. The results from our proposed method greatly exceed the second team in the different evaluation metrics. Compared with the method from the second team, our method gains a 0.126 increase of Macro F1, 0.141 increase of Macro Precision, 0.124 increase of Macro Recall, and 0.108 increase of Micro F1, Micro Precision, Micro Precision, and Accuracy. The proposed method obtains the state-of-the-art performance from the perspective of emotion classification and achieves substantial improvements over other methods.

As for the ablation study part, we implement different ablation settings to show the effectiveness of the proposed method. As shown in Table 3, the PLM model contributes a lot for the emotional classification. The continuing pre-training can further improve the emotion classification on the three metrics based on the original pre-trained language model. Other experimental results also demonstrate that the training strategies are important for better results. More concretely, the proposed supervised transfer, child-tuning, and late fusion methods help improve the final results.

5 Conclusion

This paper illustrates our contributions to the WASSA shared work on Emotion Classification. We use the DeBERTa pre-trained language model enhanced by the continual pre-training method (MLM) and some training strategies to improve the EMO performance. During the evaluation phase, our submission achieves Top-1 on all metrics for the Emotion Classification task. In the future, we will explore more efficient pre-training methods to

further improve the final results.

Acknowledgement

This work is supported by the National Key R&D Program of China (2018YFB1305200), the National Natural Science Fund of China (62171183).

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *arXiv: Computation and Language*.

Yash Butala, Kanishk Singh, Adarsh Kumar, and Shrey Shrivastava. 2021. Team phoenix at wassa 2021: Emotion analysis on news stories with pre-trained language models. *arXiv: Computation and Language*.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Sean X Chen and Jun S Liu. 1997. Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, pages 875–892.

Niko Colnerič and Janez Demšar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*, 11(3):433–446.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep nlp models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Tommaso Fornaciari, Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. Milanlp @ wassa: Does bert feel sad when you cry?

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7780–7788.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained bert encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning*.

Atharva Kulkarni, Sunanda Somwase, Shivam Rajput, and Manisha Marathe. 2021. Pvg at wassa 2021: A multi-input, multi-task, transformer-based architecture for empathy and distress prediction. *arXiv preprint arXiv:2103.03296*.

Bing Liu. 2015. Sentiment analysis: Mining opinions, sentiments, and emotions.

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*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Linkai Luo and Yue Wang. 2019. Emotionx-hsu: Adopting pre-trained bert for emotion classification. *arXiv preprint arXiv:1907.09669*.

Jay Mundra, Rohan Gupta, and Sagnik Mukherjee. 2021. Wassa@iitk at wassa 2021: Multi-task learning and transformer finetuning for emotion classification and empathy prediction. *arXiv: Computation and Language*.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):1–19.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Zahra Rajabi, Amarda Shehu, and Ozlem Uzuner. 2020. A multi-channel bilstm-cnn model for multilabel emotion classification of informal text. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 303–306. IEEE.

Meena Rambocas and Barney G Pacheco. 2018. Online sentiment analysis in marketing research: a review. *Journal of Research in Interactive Marketing*.

Li Ran, Lin Zheng, Lin Hailun, Wang Weiping, and Meng Dan. 2018. Text emotion analysis: A survey. *Journal of Computer Research and Development*, 55(1):30.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*.