NCUEE-NLP at WASSA 2023 Empathy, Emotion, and Personality Shared Task: Perceived Intensity Prediction Using Sentiment-Enhanced RoBERTa Transformers

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee
Department of Electrical Engineering
National Central University
No. 300, Zongda Rd., Zhongli Dist., Taoyuan City 32001, Taiwan
110521087@cc.ncu.edu.tw, 111521100@cc.ncu.edu.tw, lhlee@ee.ncu.edu.tw

Abstract

This paper describes our proposed system design for the WASSA 2023 shared task 1. We propose a unified architecture of ensemble neural networks to integrate the original RoBERTa transformer with two sentiment-enhanced RoBERTa-Twitter and EmoBERTa models. For Track 1 at the speech-turn level, our best submission achieved an average Pearson correlation score of 0.7236, ranking fourth for empathy, emotion polarity and emotion intensity prediction. For Track 2 at the essay-level, our best submission obtained an average Pearson correlation score of 0.4178 for predicting empathy and distress scores, ranked first among all nine submissions.

1 Introduction

Empathy is the capacity to understand or feel what another person is experiencing from his/her perspectives, which is a cognitive and emotional reaction to observing the situation of others (Omitaomu et al., 2022). Computational detection and prediction of empathy has attracted considerable attention in recent years. Empathy assessment by the writer of a statement was captured and annotated to computationally distinguish between multiple forms of empathy, empathic concerns and personal distress (Buechel et al., 2018). Mixed-Level Feed Forward Network (MLFFN) was proposed to learn word ratings for empathy and distress (Sedoc et al., 2020). Logistic regression models were used to recognize distress and condolences reactions to such distress (Zhou and Jurgens, 2020). A multi-task RoBERTa-based bi-encoder model was developed to identify empathy in conversations (Sharma et al., 2020). A demographic-aware EmpathBERT architecture was presented to infuse demographic information for empathy prediction (Guda et al., 2021).

The Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) organizes shared tasks for different aspects of affect computation from texts. In WASSA 2021 and 2022 shared tasks that focus on predicting empathy and emotion in reaction to news stories (Barriere et al., 2022; Tafreshi et al., 2021), using transformers based pre-trained language models to achieve promising results. The PVG team proposed a multi-input and multi-task framework based on the RoBERTa transformer for empathy score prediction (Kulkarni et al., 2021). An ensemble of the RoBERTa multi-task model and the vanilla ELECTRA model was used to predict empathy scores (Mundra et al., 2021). The IUCL system fine-tuned two RoBERTa large models, including a regression model for empathy and distress prediction and a classification model for emotion detection (Chen et al. 2022). A multi-output regression model fine-tuned by RoBERTa with additional features, including gender, income and age was used to predict empathy and distress intensity (Arco et al., 2022). The task adapters for a RoBERTa model were trained to predict empathy and distress scores at the essay-level (Lahnala et al., 2022).

WASSA-2023 organizes a similar task with a newly added track on empathy, emotion and self-disclosure detection in conversation at the speech-turn level (Barriere et al., 2023). We participated in the Track 1 for Empathy and Emotion Prediction in Conversations (CONV), aiming to predict perceived empathy, emotion polarity and emotion

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 548–552

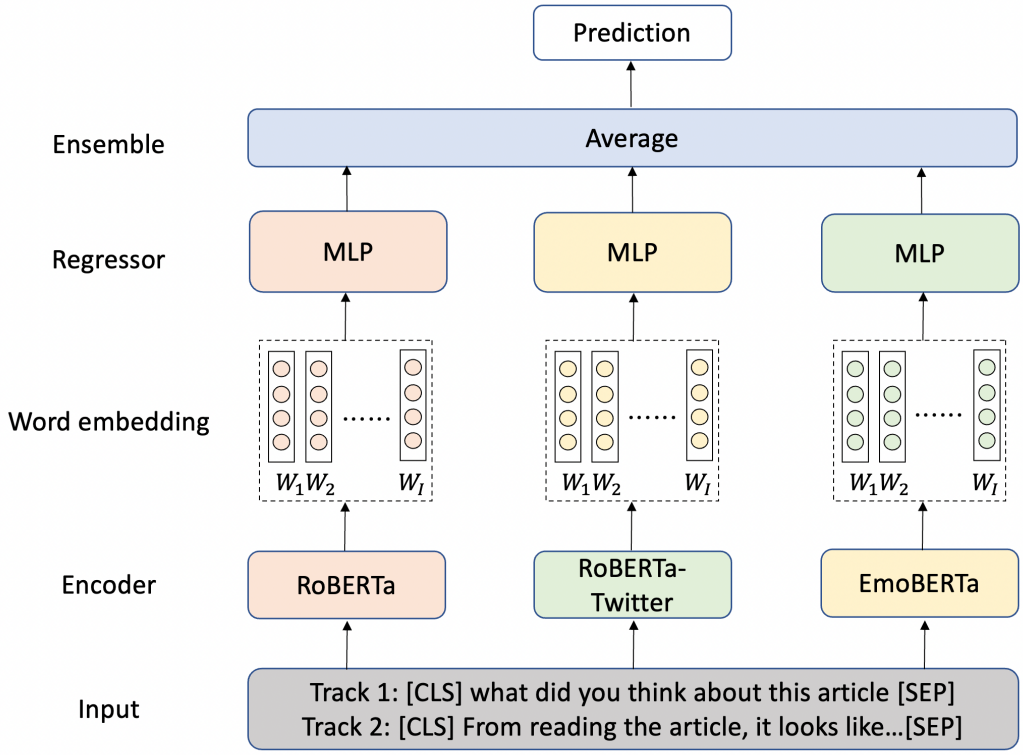

Figure 1: Our NCUEE-NLP system architecture for the Tracks 1 and 2 in the WASSA 2023 task 1.

intensity at the speech-turn-level in a conversation, and Track 2 for Empathy Prediction (EMP) to predict empathy concerns and personal distresses at the essay-level. Both tracks are regression tasks evaluated based on the average of the Pearson correlations. Following the successes of RoBERTa-based models in the previous WASSA shared tasks, we explore the use of sentiment-enhanced RoBERTa models to address the challenges for both tracks in the shared task 1.

This paper describes the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing Lab) system for WASSA 2023 shared task 1. A unified framework is used to integrate the original RoBERTa transformer (Liu et al., 2019) with different sentiment-enhanced versions, including RoBERTa-Twitter (Barbieri et l., 2020) and EmoBERTa (Kim and Vossen, 2021) for both tracks. For Track 1, our best submission achieved an average Pearson correlation of 0.7236 and ranked fourth among all participating teams. For Track 2, our best result had an average Pearson correlation of 0.4178, ranking first among all nine submissions.

The rest of this paper is organized as follows. Section 2 describes the NCUEE-NLP system for Tracks 1 and 2 in the WASSA 2023 shared task 1. Section 3 presents the results and performance comparisons. Conclusions are finally drawn in Section 4.

2 The NCUEE-NLP System

We propose a unified architecture of ensemble neural networks to solve Tracks 1 and 2 of WASSA-2023 shared task 1. Figure 1 shows our system architecture for empathy and emotion prediction, which mainly depends on ensemble sentiment-enhanced transformers. We select the following RoBERTa-based transformers to tackle both tracks in this task.

(1) Robust optimized BERT pre-training approach (RoBERTa) (Liu et al., 2019)

RoBERTa is a replication study of BERT (Devlin et al., 2019) pre-training that carefully measures the impact of key parameters and training data size. The model modifications include removing the next sentence predictions,

RoBERTa Transformers	Track 1: CONV at the speech-turn level				Track 2: EMP at the essay-level		
	Empathy	Emotion Polarity	Emotion Intensity	Average	Empathy	Distress	Average
RoBERTa	0.7715	0.7608	0.6941	0.7421	0.6660	0.5596	0.6128
RoBERTa-twitter	0.7871	0.7671	0.7061	0.7534	0.6000	0.5564	0.5782
EmoBERTa	0.7693	0.7659	0.6899	0.7417	0.6278	0.5454	0.5866
Ensemble	**0.7901**	**0.7751**	**0.7076**	**0.7576**	**0.6702**	**0.5905**	**0.6304**

Table 1: Submission results on the validation set.

dynamically changing the masking pattern applied to the training data, and training with large batches.

(2) RoBERTa-Twitter (Barbieri et al., 2020)

RoBERTa-Twitter is a RoBERTa model trained on nearly 58M tweets and fine-tuned for sentiment analysis using the TweetEval benchmark.

(3) EmoBERTa (Kim and Vossen, 2021)

EmoBERT is a RoBERTa model trained to solve emotion recognition in conversation tasks. EmoBERTa can learn speaker-aware states and contexts to predict the emotion of a current speaker by simply prepending speaker names to utterances and inserting separation tokens between the utterances in a dialogue.

For both tracks in this shared task, we fine-tune these pre-trained RoBERTa-based transformers using the datasets provided by task organizers. For Track 1 on empathy and emotion prediction in conversations, we separately fine-tuned these transformers for empathy, emotion polarity and emotion intensity prediction. For Track 2 on empathy prediction at the essay-level, we respectively trained the transformers for empathy and distress score prediction.

Finally, we use the average ensemble mechanism to combine these individual sentiment-enhanced RoBERTa transformer to produce a desired score output for both tracks.

3 Evaluation

3.1 Datasets

The experimental datasets were provided by task organizers (Barriere et al., 2023). During system development phase, the training and validation sets respectively consisted of 8,776 and 2,400 conversations for Track 1. In addition, the training and validation sets for Track 2 respectively feature 792 and 208 essays. During the evaluation period, the test sets contain 1425 conversations for Track 1 and 100 essays for Track 2.

3.2 Settings

The pre-trained RoBERTa transformers models were download from HuggingFace[1]. The hyper-parameter values for our model implementation were used as follows: epoch 25, batch size 8, learning rate 1e-5, and max sequence 256.

To confirm the average ensemble performance, we also compared individual transformers. The evaluation metric is the Pearson correlation for both tracks. For Track 1, we obtained Pearson correlation coefficients of the empathy, emotion polarity and emotion intensity at the speech-turn level. For Track 2, we had Pearson correlation coefficients for empathy and distress at the essay-level. The official ranking of each participating team was based on the average of the obtained Pearson correlation coefficients.

3.3 Results

Table 1 shows the results on the validation set. For Track 1 at the speech-turn level, RoBERTa-Twitter outperformed the other standalone transformer models for all evaluation metrics, but relatively underperformed for Track 2 at the essay level. The ensemble transformers clearly achieved the best

[1] https://huggingface.co/roberta-base
https://huggingface.co/tae898/emoberta-base
https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

RoBERTa Transformers	Track 1: CONV at the speech-turn level				Track 2: EMP at the essay-level		
	Empathy	Emotion Polarity	Emotion Intensity	Average	Empathy	Distress	Average
RoBERTa	0.7849	0.6851	0.6384	0.7028	0.3327	0.3819	0.3573
RoBERTa-twitter	0.7898	0.6941	**0.676**	0.72	0.3661	0.3415	0.3538
EmoBERTa	0.772	0.6638	0.6418	0.6925	0.4074	**0.4663**	**0.4368**
Ensemble (Official submission)	**0.8035**	**0.6981**	0.6692	**0.7236**	**0.415**	0.4206	0.4178

Table 2: Submission results on the test set.

performance for both tracks on the validation set. This confirms that the ensemble averaging mechanism works well in integrating multiple models to obtain performance improvement.

Table 2 shows the results on the test set for both tracks. For CONV Track 1, RoBERTa-twitter outperformed the others in the emotion intensity evaluation at the speech-turn level. Our ensemble sentiment-enhanced RoBERTa model achieved the best average Pearson correlation coefficient of 0.7236. For EMP Track 2, EmoBERTa obtained the best distress and average correlation coefficients, while our ensemble transformer model achieved the second-best correlation coefficient of 0.4178.

3.4 Rankings

According to official rankings released by task organizers (Barriere et al., 2023), our final submission from ensemble neural networks of sentiment-enhanced RoBERTa transformers ranked fourth for Track 1 and first for Track 2 among all nine submissions.

4 Conclusions

This study describes the model design, system implementation and performance of the NCUEE-NLP system in the WASSA 2023 Task 1 for empathy and emotion prediction. Our unified architecture used an average ensemble mechanism of three sentiment-enhanced RoBERTa transformers to predict empathy, emotion polarity and emotion intensity for Track 1 at the speech-turn level and empathy and distress scores for Track 2 at the essay-level. Our final submission based on sentiment-enhanced RoBERTa transformers ranked fourth for Track 1 and first for Track 2 among all nine submissions.

Acknowledgments

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

References

Flor Miriam Del Arco, Jaime Collado-Montañez, L. Alfonso Ureña, and María-Teresa Martín-Valdivia. 2022. Empathy and distress prediction using transformer multi-output regression and emotion analysis with an ensemble of supervised and zero-shot learning models. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Association for Computational Linguistics, pages 239–244. https://doi.org/10.18653/v1/2022.wassa-1.23

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pages 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, pages 214-227. http://dx.doi.org/10.18653/v1/2022.wassa-1.20

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. WASSA 2023 shared task: predicting empathy, emotion and personality in reactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 4758–4765. https://doi.org/10.18653/v1/D18-1507.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022. IUCL at WASSA 2022 shared task: a text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Association for Computational Linguistics, pages 228–232. https://doi.org/10.18653/v1/2022.wassa-1.21

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)*. Association for Computational Linguistics, pages 4171-4186. http://dx.doi.org/10.18653/v1/N19-1423

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: a BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, pages 3072–3079. https://doi.org/10.18653/v1/2021.eacl-main.268.

Taewoon Kim and Piek Vossen. 2021. EmoBERa: speaker-aware emotion recognition in connersation with RoBERTa. *arXiv preprint,* https://arxiv.org/abs/2108.12009.

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. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, pages 105–111.

Allison Lahnala, Charles Welch, and Lucie Flek. 2022. CAISA at WASSA 2022: adapter-tuning for empathy prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Association for Computational Linguistics, pages 280–285. https://doi.org/10.18653/v1/2022.wassa-1.31

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.

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. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, pages 112–116.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. Empathic conversations: a multi-level dataset of contextualized conversations. arXiv Preprint, arXiv:2205.12698. https://doi.org/10.48550/arXiv.2205.12698

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, pages 1664–1673.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 5263–5276. https://doi.org/10.18653/v1/2020.emnlp-main.425.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, Alexandra Balahur. 2021. WASSA 2021 shared task: predicting empathy and emotion in reaction to news stories. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 92–104.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 609–626. https://doi.org/10.18653/v1/2020.emnlp-main.45