NYCU-NLP at EXALT 2024: Assembling Large Language Models for Cross-Lingual Emotion and Trigger Detection

Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, and Lung-Hao Lee*
Institute of Artificial Intelligence Innovation
National Yang Ming Chiao Tung University
No. 1001, Daxue Rd., East Dist., Hsinchu City 300093, Taiwan
*lhlee@nycu.edu.tw

Abstract

This study describes the model design of the NYCU-NLP system for the EXALT shared task at the WASSA 2024 workshop. We instruction-tune several large language models and then assemble various model combinations as our main system architecture for cross-lingual emotion and trigger detection in tweets. Experimental results showed that our best performing submission is an assembly of the Starling (7B) and Llama 3 (8B) models. Our submission was ranked sixth of 17 participating systems for the emotion detection subtask, and fifth of 7 systems for the binary trigger detection subtask.

1 Introduction

Emotion detection is a well-studied NLP task that aims to automatically identify affective information from texts. The EXALT task organized within the WASSA-2024 workshop focuses on the explainability of cross-lingual emotion detection in tweets. In the emotion detection subtask, the participating system should predict for each tweet an emotion label from 6 possible classes: Love, Joy, Anger, Fear, Sadness, and Neutral. To investigate transferable emotion information across languages, training data is provided in English and evaluation data consists of five different target languages: Dutch, Russian, Spanish, French and English. In the binary trigger detection subtask, participating systems should further identify which words or emoticons can be used to express the emotion.

This paper describes the **NYCU-NLP** (**N**ational **Y**ang Ming **C**hiao Tung **U**niversity, **N**atural **L**anguage **P**rocessing Lab) system for the EXALT shared task (Maladry et al., 2024). Our solution explores the use of instruction-tuned LLMs, including Mistral (7B) (Jiang et al., 2023), Starling (7B) (Zhu, 2023) and Llama 3 (8B) (Meta AI, 2024). We then assemble various model combinations as our main system architecture. Experimentally, our best performing submission was an assembly of Starling (7B) and Llama 3 (8B), which was ranked sixth of 17 participating systems for the emotion detection subtask and fifth of 7 systems for the binary trigger detection subtask.

The rest of this paper is organized as follows. Section 2 reviews recently related studies on emotion detection. Section 3 describes the NYCU-NLP system for the EXALT shared task. Section 4 presents the results and performance comparisons. Conclusions are finally drawn in Section 5.

2 Related Work

Transformer-based language models have been widely applied to emotion detection. An ensemble of the BERT and ELECTRA models was used to detect emotions (Kane et al., 2022). A knowledge-enriched transformer was designed for emotion detection in textual conversations (Zhong et al., 2019). A topic-driven transformer was proposed to detect emotions within dialogues (Zhu et al., 2021). Two hierarchical transformers were trained to use context-/speaker-sensitive information for emotion detection in conversations (Li et al., 2020). Sentiment-enhanced RoBERTa transformers were used to predict emotion and empathy intensities (Lin et al., 2023). Empirical evaluations also showed that transformer-based models such as BERT and XLNet outperformed conventional neural networks for sentiment intensity prediction (Lee et al., 2022). A transformer-based fusion model was developed to integrate semantic

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 505–510

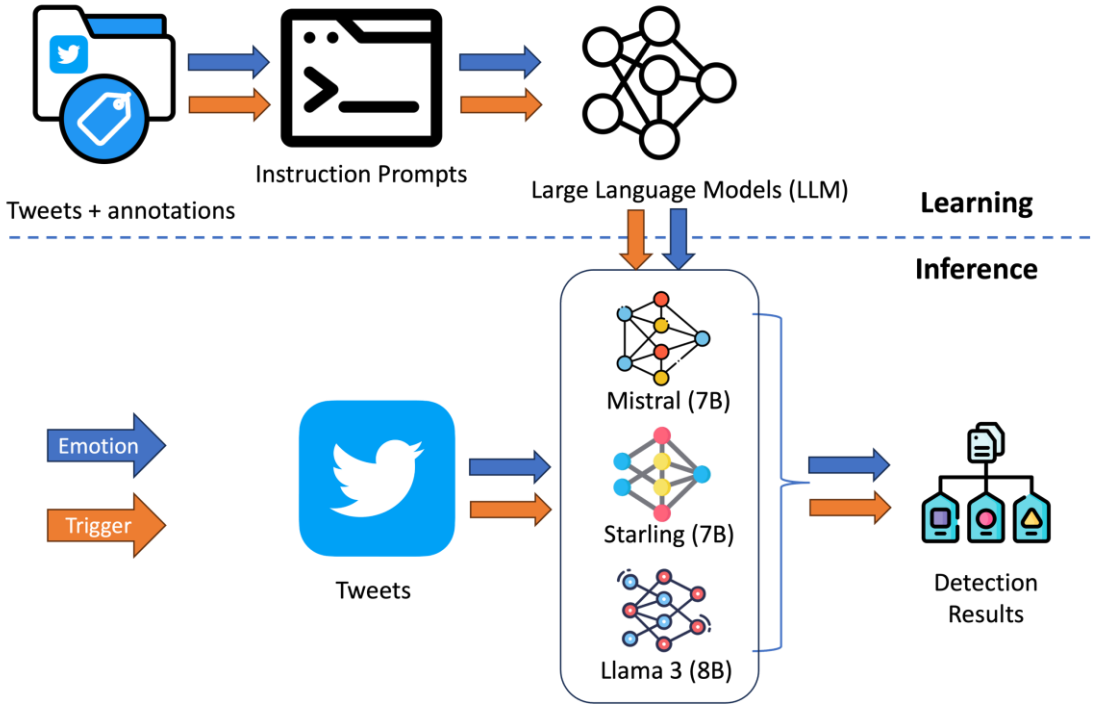

Figure 1: Our NYCU-NLP system architecture for the EXALT shared task.

representations at different degrees of linguistic granularity for emotional intensity prediction (Deng et al., 2023).

Recently, Large Language Models (LLM) have been used for emotion detection. Fine-tuned GPT-3 models with prompt engineering for zero-shot or few shot learning with ChatGPT and GPT-4 models were evaluated for emotion detection (Nedilko and Chu, 2023). Prompt engineering techniques were applied to a GPT model for emotion detection in a code-switching setting (Nedilko, 2023). Multiple features generated by ChatGPT were integrated for emotion recognition in conversations (Tu et al., 2023). The LLM-GEm system was designed to use GPT 3.5 for empathy prediction (Hasan et al., 2024). The abilities of CPT-4, Llama2-Chat-13B and Aplaca-13B to identify emotion triggers were evaluated, analyzing the importance of trigger words for emotion prediction (Singh et al., 2024). Given the results obtained by most such approaches, we are motivated to explore the application of LLMs to the emotion and trigger detection tasks.

3 The NYCU-NLP System

Figure 1 shows our NYCU-NLP system architecture for the EXALT shared task. We first instruction-tune LLMs and then assemble fine-tuned LLMs for cross-lingual emotion and binary trigger detection in tweets.

3.1 Large Language Models

The following LLMs are used to explore the effectiveness of our system architecture.

(1) Mistral (7B) (Jiang et al. 2023)

Mistral-7B is an open source LLM under the Apache 2.0 license which leverages Group-Query Attention (CQA) for faster inference and uses Sliding Window Attention (SWA) to handle longer sequences at smaller cost. Mistral-7B claims it outperforms Llama 2 (13B) across all evaluated benchmarks.

(2) Starling (7B) (Zhu et al., 2023)

Starling-7B is an open LLM trained by Reinforcement Learning from AI Feedback (RLAIF). A new ranking dataset, called Nectar, was used for the proposed new reward training and policy tuning pipeline. Starling-7B was mainly evaluated based on MT-Bench and AlpacaEval, which are GPT-4-based comparisons.

(3) Llama 3 (8B) (Meta AI, 2024)

Llama 3 is Meta's next generation release of the well-known Llama model. We use a pretrained and instruction-fine-tuned Llama 3 model with 8B parameters.

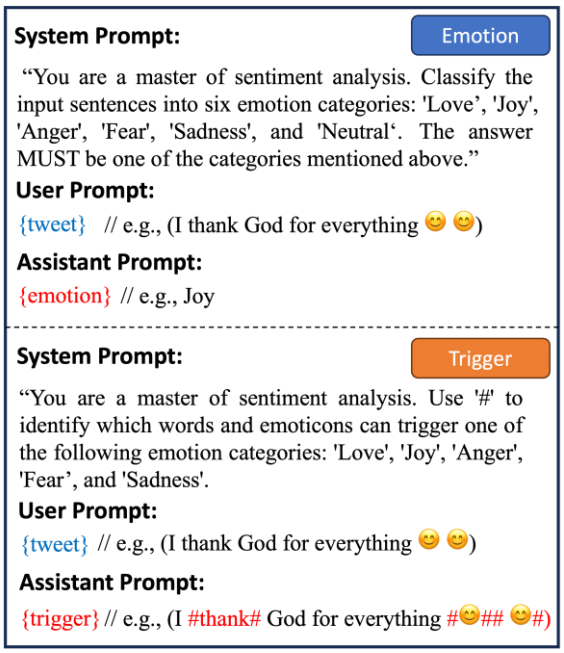

Figure 2: Prompts used for instruction fine-tuning.

3.2 Instruction Fine-tuning

We continually fine-tune these three publicly released LLM models. Figure 2 shows the prompts used for instruction fine-tuning (Wei et al., 2022). The system is configured as a master of sentiment analysis for both tasks. For the emotion detection subtask, we ask the LLM to classify the given sentence into six defined emotion categories. For the binary trigger detection subtask, we used the "#" symbol to emphasize words and emoticons that can be used to trigger the emotion.

We also use the Low-Rank Adaption (LoRA) technique (Hu et al., 2021), which freezes the pre-trained LLM weights and injects trainable rank decomposition matrices into each layer of the transformer architecture, greatly facilitating the instruction-tuning process for downstream tasks.

3.3 Assembly Mechanism

During the inference phase, we use a voting-based assembly mechanism, which each LLM conducting an independent detection for each testing instance, effectively integrating fine-tuned LLMs to determine the system output by a majority of votes.

For the emotion detection subtask, if a testing instance does not have a major category prediction, we use the 'neutral' emotion category as an alternative option.

For cases without a majority of prediction results in the binary trigger detection subtask, if a word or emoticon is identified by any one of our used models, we directly regard it as a trigger for our system output.

4 Performance Evaluation

4.1 Data

The datasets were mainly provided by task organizers, including the training data in English, along with development and test data for each of the five target languages. For the emotion detection subtask, there are respectively 5000, 3000 and 2500 instances in the training, development and test sets, and we used these datasets without augmentation. For the binary trigger subtask, we had respectively 3000, 300 and 832 instances for each provided dataset.

4.2 Settings

All pretrained models were downloaded from HuggingFace [1]. We continuously fine-tuned the LLM models using the training datasets only. All experiments were conducted on a server with two Nvidia V100 GPUs (Total 64GB memory). The hyperparameter values for our model implementation were manually optimized on the given development set as follows: epochs 20; batch size 2; optimizer AdamW; learning rate 1e-4; LoRA r 16; LoRA alpha 32; LoRA drop 0.1 and max token length of 20.

4.3 Results

Table 1 shows the submission results on the development set. Among three independent LLMs, Llama 3 (8B) outperformed the others in the terms of F1-scores on both subtasks. Assemble LLMs usually outperformed independent LLMs, except the Mistral may reduce performance. The best performance was achieved by an assembly of Starling (7B) and Llama 3 (8B), so we use this LLM setting either individually or in combination as our final submission for official ranking.

Table 2 shows the submission results on the test set. Independent Llama 3 (8B) outperformed independent Starling (7B) for all evaluation metrics.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
https://huggingface.co/Nexusflow/Starling-LM-7B-beta
https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

LLMs (#para)	Emotion Detection			Binary Trigger Detection			
	Prec.	Recall	F1	Token Prec.	Token Recall	Token F1	MAP
Mistral (7B)	0.5491	0.5114	0.5202	0.4239	0.4094	0.3768	0.3697
Starling (7B)	0.5677	0.4909	0.5127	0.6321	0.5430	0.5442	0.5219
Llama 3 (8B)	0.5394	0.5713	0.5784	**0.6857**	0.5641	0.5701	0.5438
Mistral + Starling	0.5771	0.5590	0.5641	0.6316	0.4906	0.5126	0.5058
Mistral + Llama 3	0.5894	0.5728	0.5793	0.6774	0.5461	0.5605	0.5398
Starling + Llama 3	**0.6383**	**0.5778**	**0.5982**	0.6277	**0.6597**	**0.5836**	0.5156
Mistral + Starling + Llama 3	0.6376	0.5617	0.5866	0.6849	0.5798	0.5770	**0.5466**

Table 1: Submission results on the development set.

LLMs (#para)	Emotion Detection			Binary Trigger Detection			
	Prec.	Recall	F1	Token Prec.	Token Recall	Token F1	MAP
Starling (7B)	0.5636	0.5416	0.5496	0.5946	0.4454	0.4673	0.4649
Llama 3 (8B)	0.5872	**0.5806**	0.5815	**0.6601**	0.4859	0.5179	0.5103
Starling + Llama 3	**0.6200**	0.5788	**0.5951**	0.6442	**0.5901**	**0.5636**	**0.5162**

Table 2: Submission results on the test set.

The assembly of Starling and Llama 3 obtained the best F1 score of 0.5951 for the emotion detection subtask, ranking the sixth of 17 participating systems. In addition, this assembly achieved the best token F1 of 0.5636 for the binary trigger detection subtask, ranking the fifth among all 7 participating systems.

4.4 Discussion

We did not use prompt engineering techniques to configure other prompts due to limited computational resources. Therefore, prompts used for instruction fine-tuning needed to be improved for performance enhancements.

In addition, since the LLMs were pre-trained using multi-lingual data, we do not use any machine translation techniques in the tasks.

5 Conclusions

This study describes the NYCU-NLP system for the EXALT shared task at the WASSA 2024 workshop, including model design and performance evaluation. We instruction-fine-tuned the LLMs to effectively detect cross-lingual emotions and triggers. Experimental results indicate that our best submission is an assembly of Starling (7B) and Llama 3 (8B) models, achieving a F1 score of 0.5961 for the emotion detection subtask (ranking sixth out of seventeen) and a token F1 of 0.5636 for the binary trigger detection subtask (ranking fifth out of seven).

This pilot study is our first exploration in the cross-lingual emotion and trigger detection task. In future, we will exploit other advanced LLMs to further improve performance.

Acknowledgments

This study is partially supported by the National Science and Technology Council, Taiwan, under the grant NSTC 111-2628-E-A49-029-MY3. This work was also financially supported by the Co-creation Platform of the Industry Academia Innovation School, NYCU.

Limitations

This work does not propose a new model to address this shared task. Due to computational resource limitations, experiments were conducted

with basic settings without other advanced explorations to enhance system performance.

References

Meta AI (2024). Llama 3 (April 18 version) [Large language model]. https://ai.meta.com/blog/meta-llama-3/

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Towards transformer fusions for Chinese sentiment intensity prediction in valence-arousal dimensions. *IEEE Access*, 11:109974-109982. https://doi.org/10.1109/ACCESS.2023.3322436

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics, pages 2214-2231.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint*, *arXiv:2106.09685v2*. https://doi.org/10.48550/arXiv.2106.09685

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lelio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothee, Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv: 2310.06825v1*. https://doi.org/10.48550/arXiv.2310.06825

Aditya Kane, Shantanu Patankar, Sahil Khose, and Neeraja Kirtane, 2022. Transformer based ensemble for emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics, pages 250-254. https://doi.org/10.18653/v1/2022.wassa-1.25

Lung-Hao Lee, Jian-Hong Li, ad Liang-Chih Yu. 2022. Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4): Article 65, 1-18. https://doi.org/10.1145/3489141

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, pages 4190-4200. https://doi.org/10.18653/v1/2020.coling-main.370

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 Empathy, Emotion, and Personality Shared Task: Perceived intensity prediction using sentiment-enhanced RoBERTa transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics, pages 548-552. https://doi.org/10.18653/v1/2023.wassa-1.49

Andrew Nedilko. 2023. Generative pretrained transformers for emotion detection in a code-switching setting. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics, pages 616-620. https://doi.org/10.18653/v1/2023.wassa-1.61

Andrew Nedilko, and Yi Chu. 2023. Team Bias Busters at WASSA 2023 Empathy, Emotion, and Personality shared task: emotion detection with generative pretrained transformers. n *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics, pages 569-573. https://doi.org/10.18653/v1/2023.wassa-1.53

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2024. Language models (mostly) do not consider emotion triggers when predicting emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 603-614.

Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. 2023. An empirical study on multiple knowledge from ChatGPT for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, pages 12160-12173. https://doi.org/10.18653/v1/2023.findings-emnlp.813

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *Proceedings of the 10th International Conference on Learning Representations*. arXiv:2109.01652v5. https://doi.org/10.48550/arXiv.2109.01652

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Strling-7B: Increasing LLM

helpfulness & harmlessness with RLAIF. https://starling.cs.berkeley.edu/

Peixiang Zhong, Di Wang, Chunyan Miao. 2019. Knowledge-enriched Transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, pages 165-176. https://doi.org/10.18653/v1/D19-1016

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1571-1582. https://doi.org/10.18653/v1/2021.acl-long.125