Bridging the Gap in Text-Based Emotion Detection

Sarvnazsadat Roumianfar
Ali Abbasi
Politecnico di Torino

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Bridging the Gap in Text-Based Emotion Detection

***Abstract*—This study presents a fine-tuned large language model for multi-label emotion classification based on the BRIGHTER dataset [1], as part of the SemEval 2025 Task 11, Track A [2]. The purpose was to develop a transformer-based classifier capable of detecting multiple emotions simultaneously from short social media posts. The abstract does not only mention the paper, but is the original paper shrunken to approximately 200 words. It states the purpose, reports the information obtained, gives conclusions, and recommendations. In short, it summarizes the main points of the study adequately and accurately. It provides information from every major section in the body of the report in a dense and compact way. Past tense and active voice is appropriate when describing what was done. If there is any, it includes key statistical detail.**

Depending on the format you use, the abstract may come on the title page or at the beginning of the main report.

I. Introduction

Emotion detection from text is a core NLP task with applications in areas like social media monitoring and mental health analysis. Unlike single-label sentiment classification, multi-label emotion classification captures the co-occurrence of multiple emotional states in one text, making the task more challenging. We use the BRIGHTER dataset [1], [2], a multilingual benchmark for this task, which includes emotion-annotated texts in 28 languages—focusing especially on low-resource languages across Africa, Asia, Eastern Europe, and Latin America. The data is multilabeled with five core emotion categories: joy, sadness, fear, anger, and surprise. Each sample was annotated by native speakers to ensure linguistic quality and emotional relevance. Our work uses the English subset of BRIGHTER for training and evaluation.

II. Problem Definition–sarv

The problem is to design a model capable of detecting multiple concurrent emotional states from short text inputs. The challenge includes:

- Effectively handling multi-label outputs: Each input sentence from BRIGHTER dataset [1] can include one or more emotions simultaneously, so the model should be able to detect and predict multiple emotions.
- Working with imbalanced labeled data: Some emotions like fear and sadness are more frequent in the dataset [1] than other feelings like anger, which makes it harder for model to recognize and detect rare emotions.
- Training efficiently under compute constraints: Full fine-tuning of large language models is resource-intensive. Therefore, we had to apply some techniques like quantization [4] to reduce model size and memory usage and parameter fine-tuning methods like LoRA [3] to train effectivly with limited hardware resources.

Fig. 1. Simulation Results

III. Proposed Solutions–ali and sarvnaz

A. Transformer-Based Model Fine-Tuning

We fine-tuned pre-trained transformer models, specifically BERT and DeBERTa, LLama to perform multi-label classification.

B. Loss Function Adjustment with Class Weights

To handle the effect of imbalanced labels, we computed class weights for each emotion label in the training data, so we managed to improve the model's sensitivity to underrepresented emotions.

C. Layer Unfreezing

To balance performance with computational efficiency, we experimented with freezing the early layers of the transformer model and fine-tuning only the final layers. we have have tried unfreezing different layers of llma trasnformer. The results showed that unfreezing last 3 (???????) layer has brought us the most effective and most accuracte one. ??????????

D. Lightweight Fine-Tuning with LoRA and Qunatization

In order to train under limited computational resources, we applied Low-Rank Adaptation (LoRA) to introduce task-specific trainable parameters while keeping the majority of the model frozen. Additionally, we employed **4-bit quantization** to compress the model weights and reduce memory consumption in training.

IV. Criteria for Assessing Solutions- sarvnaz

F1-Score (Micro and Macro) [7]: Micro F1 calculates metrics globally by counting total true positives, false negatives, and false positives, reflecting overall model performance. Macro F1 computes the metric independently for each class and then averages them, treating all classes equally regardless of their frequency.

V. Research Methodology- ali

We used the English subset of the BRIGHTER dataset for training, validation, and testing. Each text sample was multi-labeled and were transformed into binary vectors of length five, corresponding to the five core emotions.

During experimentation, we implemented class frequency analysis to compute emotion-wise weights for the loss function. We froze the transformer backbone initially, then selectively unfroze the final layers to fine-tune only a subset of parameters. To evaluate model performance, we used epoch-based validation and computed macro and micro F1 scores to account for both frequent and rare emotions.

——————————— tozihat

The main difference between this section and the one in your report proposal is use of verb tense: there you suggested what you will do and here you will describe what you did. Be concise and precise when outlining how you researched your potential solutions. Remember that your research should be guided by:

- Relevance to the context of application
- Your assessment criteria
- Practicality

So it may be worth commenting on your research methodology in light of the above (e.g., justifying a particular approach).

In this section, only describe how you collected data, and explain what you did to test your criteria. *Do not include your findings in this section.* ———————————

VI. Analysis and Interpretation- sarv -ali

In this section you will mainly analyze your data in terms of your assessment criteria; e.g., do the data suggest that a particular solution is "cost effective" "environmentally acceptable", "technically feasible" or "affordable"?

Be logical and selective when analyzing/interpreting your research data. For example, if a proposed solution is proven to be far too expensive to realistically implement in your context, is there any value in discussing whether it is "culturally viable" or "technically sustainable"? Perhaps in this case you can focus more attention on solutions that your research suggests are more valid. Do not just throw huge quantities of raw data at your reader and leave them to interpret it. Present enough to transparently support any conclusions you draw and make sure that you offer justifications for your analysis.

Be honest and reflective while discussing your data. Your data might be too limited or unclear to interpret with accuracy—explain this, perhaps suggesting how this shortcoming could be addressed. Admitting the above will help you draw more honest and worthwhile conclusions.

Remember that research is an imperfect and ongoing process that should be open to question and verification. Therefore, unless convinced by the absolute strength of your evidence, you should be tentative in your language choice when interpreting/analyzing research results. Selectively use *hedging* (language which indicates a lack of certainty) to modify the tone of your analysis and any conclusions that result from this.

Here are some examples that show differing degrees of certainty:

	Growth Media				
model	1	2	3	4	5
meta-llama/Llama-3.2-1B	0.962	0.821	0.356	0.682	0.801
NWN652	0.981	0.891	0.527	0.574	0.984
PPD234	0.915	0.936	0.491	0.276	0.965
JSB126	0.828	0.827	0.528	0.518	0.926
JSB724	0.916	0.933	0.482	0.644	0.937
Average Rate	0.920	0.882	0.477	0.539	0.923

TABLE I. Performance Comparison of Transformer-Based Emotion Classifiers

- it appears that …
- it can be tentatively concluded that …
- it is almost certain that …
- perhaps the evidence indicates …
- this seems to point to the fact that …
- this could be interpreted as evidence of …
- without doubt its application would prove beneficial for …

Finally, don't introduce any new content (e.g., research methods or solutions) within this section—this will prove confusing for the reader. The reader should clearly understand that you are, based on specific criteria, interpreting the results of your research in order to test the viability of various solutions to remedy a particular problem. The sole function of this part of the report is to openly discuss your research findings in order to set up your conclusions/recommendations.

A reference to Table I.

VII. Conclusions and Recommendations- ali

Conclusion shows what knowledge comes out of the report. As you draw a conclusion, you need to explain it in terms of the preceding discussion. You are expected to repeat the most important ideas you have presented, without copying. Adding a table/chart summarizing the results of your findings might be helpful for the reader to clearly see the most optimum solution(s).

It is likely that you will briefly describe the comparative effectiveness and suitability of your proposed solutions. Your description will logically recycle language used in your assessing criteria (section IV): "Solution A proved to be the most cost effective of the alternatives" or "Solution B, though a viable option in other contexts, was shown to lack adaptability". Do not have detailed analysis or lengthy discussions in this section, as this should have been completed in section X.

As for recommendations, you need to explain what actions the report calls for. These recommendations should be honest, logical and practical. You may suggest that one, a combination, all or none of your proposed solutions should be implemented in order to address your specific problem. You could also urge others to research the issue further, propose a plan of action or simply admit that the problem is either insoluble or has a low priority in its present state.

The recommendations should be clearly connected to the results of the report, and they should be explicitly presented. Your audience should not have to guess at what you intend to say.

APPENDIX A
WHAT GOES IN THE APPENDICES

The appendix is for material that readers only need to know if they are studying the report in depth. Relevant charts, big tables of data, large maps, graphs, etc. that were part of the research, but would distract the flow of the report should be given in the Appendices.

APPENDIX B
FORMATTING THE APPENDICES

Each appendix needs to be given a letter (A, B, C, etc.) and a title. LaTeXwill do the lettering automatically.

REFERENCES

[1] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, J. P. Wahle, T. Ruas, M. Beloucif, C. de Kock, N. Surange, D. Teodorescu, I. S. Ahmad, D. I. Adelani, A. F. Aji, F. D. M. A. Ali, I. Alimova, V. Araujo, N. Babakov, N. Baes, A.-M. Bucur, A. Bukula, G. Cao, R. Tufiño, R. Chevi, C. I. Chukwuneke, A. Ciobotaru, D. Dementieva, M. S. Gadanya, R. Geislinger, B. Gipp, O. Hourrane, O. Ignat, F. I. Lawan, R. Mabuya, R. Mahendra, V. Marivate, A. Piper, A. Panchenko, C. H. P. Ferreira, V. Protasov, S. Rutunda, M. Shrivastava, A. C. Udrea, L. D. A. Wanzare, S. Wu, F. V. Wunderlich, H. M. Zhafran, T. Zhang, Y. Zhou, and S. M. Mohammad, *BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages*. arXiv, 2025. Available: https://arxiv.org/abs/2502.11926.

[2] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, S. M. Yimam, J. P. Wahle, T. Ruas, M. Beloucif, C. De Kock, T. D. Belay, I. S. Ahmad, N. Surange, D. Teodorescu, D. I. Adelani, A. F. Aji, F. Ali, V. Araujo, A. A. Ayele, O. Ignat, A. Panchenko, Y. Zhou, and S. M. Mohammad, *SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection*. arXiv, 2025. Available: https://arxiv.org/abs/2503.07269.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv, 2021. Available: https://arxiv.org/abs/2106.09685.

[4] J. Lang, Z. Guo, and S. Huang, *A Comprehensive Study on Quantization Techniques for Large Language Models*. arXiv, 2024. Available: https://arxiv.org/abs/2411.02530.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, *LLaMA: Open and Efficient Foundation Language Models*. arXiv, 2023. Available: https://arxiv.org/abs/2302.13971.

[6] P. He, X. Liu, J. Gao, and W. Chen, *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. arXiv, 2021. Available: https://arxiv.org/abs/2006.03654.

[7] J. Opitz and S. Burst, *Macro F1 and Macro F1*. arXiv, 2021. Available: https://arxiv.org/abs/1911.03347.