Bridging the Gap in Text-Based Emotion Detection

Sarvnazsadat Roumianfar, Ali Abbasi
Politecnico di Torino, Turin, Italy

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, as a part of the SemEval 2025 Task 11, Track A. The purpose was to develop a transformer-based classifier capable of detecting multiple emotions simultaneously from short social media posts. Due to the shortage of resources that we had, we tried to adapt to solutions that need less GPU and memory resources while keeping our performance comparable to other result in the competition. Our final model achieved a macro F1 score of 0.7114 on the English subset of the BRIGHTER dataset, outperforming all leaderboard submissions to date.**

I. Introduction

Emotion classification in text is challenge due to the complexity of human expression. Accurately assigning a label to a sentence or phrase is difficult because individuals often express their emotions in abstract way or use implicit meanings. As part of the SemEval-2025 Task 11 competition [6], this project aims to develop large language model (LLM)-based architectures capable of identifying multiple emotions within English-language texts. We use the BRIGHTER dataset [5], a multilingual benchmark for this task, which includes emotion-annotated texts in 28 languages. This report outlines the evolution of the project from traditional pretrained transformers such as BERT*** and DeBERTa*** to more advanced LLMs like LLaMA***, which ultimately yielded the highest performance.

II. Problem Definition

The problem is to design a model capable of detecting multiple concurrent emotional states from short text inputs. SemEval-2025 Task 11 (Track A) defines this problem in a monolingual setting, where the goal is to predict multiple binary labels across 6 emotion categories: joy, sadness, anger, fear, surprise, and disgust (While the disgust label is not present for the English subset of dataset and they are all null). Each label is independently annotated, and the model must learn to correctly identify all applicable emotions for a given sentence. The challenge includes:

- Semantic ambiguity, where phrases can evoke different emotions based on the context.
- Class imbalance, where some emotions (e.g., fear) are overrepresented and others (e.g., anger) are rare (see Figure 1).
- Training efficiently under compute constraints, where 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.

Given these challenges, the primary objective of this project is to develop a robust and scalable LLM-based solution that accurately predicts multilabel emotional states from English text.

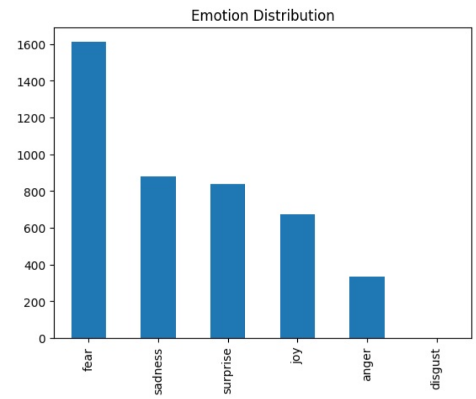

Fig. 1. Emotion distribution in the English subset of the dataset, illustrating class imbalance. Some emotions, such as fear, are significantly overrepresented, while others, such as anger are underrepresented (disgust class is absent for english dataset).

III. Proposed Solutions

A. Transformer-Based Model Fine-Tuning (encoder-only)

We explored several transformer-based models. Since, multilabel emotion Detection is a Classification problem, Encoder-Only Models are the best choice for this kind of Tasks. Our early attempts involved fully fine-tuning pretrained light models like BERT and DeBERTa

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. Transformer-Based Model Fine-Tuning (decoder-only)

Since we didn't get satisfying result from BERT and DeBERTa (Figure 2) we were looking for other models that are more powerful and larger scale. LLaMA was our choice but they were 2 problems with LLamA: First, LLaMA being a decoder-only model, is not naturally suited for classification task. However, It can be adapted for multi-label classification by adding a classification head. Second problem is that It is impossible to fully fine-tune llama due to our limited resources for this large scale model. The solution for that is freezing most of the layers of that.

D. Layer freezing

To balance performance with computational efficiency we used layer freezing. Layer freezing means preventing some layers of a model from updating during training and just fine-tune the last layers. We often freeze the first layers because they capture general, low-level features (like word or token patterns), and unfreeze the last layers to let the model learn task-specific representations without forgetting the core language knowledge. We have experimented with freezing the early layers of the model and unfreezing different layers of LLaMA. The results showed that unfreezing last 4 layers has brought us the most effective and most accurate one.

E. Lightweight Fine-Tuning with LoRA and Qunatization

In order to train under limited computational resources, we also 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.

Finally we do the hyperparameter tuning by just changing hyperparameters like batch size, number of epochs, learning rate, etc., to find the best-performing model

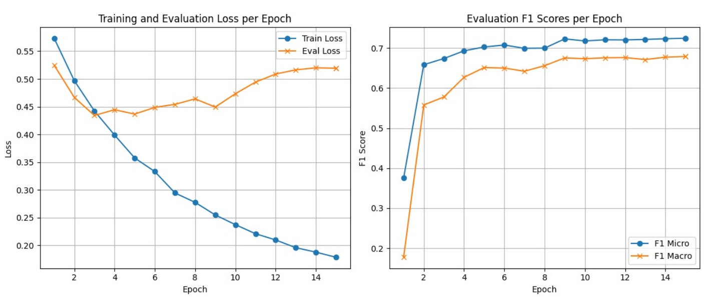

Fig. 2. Training and evaluation metrics for the DeBERTa model. While training loss steadily decreases, evaluation loss begins to increase after epoch 5, indicating overfitting. Similarly, evaluation F1 scores (both micro and macro) fluctuate rather than improve, suggesting that DeBERTa did not yield notably satisfying results on our task.

IV. Criteria for Assessing Solutions

To evaluate the performance of our models, we consider 3 metrics below:

- Micro F1-Score looks at all predictions together. It counts all correct and incorrect predictions across all classes, showing how well the model performs overall.
- Macro F1-Score treats each class equally. It calculates the F1-Score for each class separately and then averages them. This helps us to see how good the model performs on common and rare emotions.
- Training and validation loss per epoch to observer the behavior of the model in each epoch to detect Overfitting (E.g. Fig. 2), Underfitting, and the right balance between them. They can also help us to decide when to stop training or still the model can be trained without overfitting.

V. Research Methodology

We focused on building a multi-label emotion classification system using the English subset of the BRIGHTER dataset. Our methodology was based on three main considerations: (i) relevance to real-world emotion recognition tasks, (ii) alignment with our evaluation criteria (macro and micro F1-score), and (iii) practical feasibility given our computational resources.

- **Dataset:** Our experiments are based on the BRIGHTER-emotion-categories dataset, published as part of SemEval-2025 Task 11. The dataset consists of short social media posts annotated with multi-label emotion tags. Each instance can include one or more of five core emotions: joy, sadness, fear, anger, and surprise. Labels are represented using a multi-hot encoding scheme to represent the multi-label structure of the task.
  BRIGHTER includes data in 28 languages; however, For this submission, we used only the English subset ("eng"), which includes:
  - Training set: Multilabel annotations in binary format.
  - Development set: A validation split is used for model tuning and early stopping, with gold labels.
  - Test set: Used for final evaluation, with gold labels that are provided for offline scoring.
- **Data Preprocessing:** Emotion labels were coverted as binary vectors of length five to support multi-label classification. Then the dataset was tokenized using a pretrained tokenizer, and split into training, validation, and test sets.
- **Model Selection and Baselines:** All the models that we have selected, are based on the transformer architecture,that uses multi-head self-attention mechanisms, feed-forward neural networks, residual connections, and layer normalization. We chose BERT, DeBERTa, and LLaMA that are all pre-trained models, because of their great performance in different NLP tasks. They allowed us to compare encoder-based and decoder-only structures. Each model was adapted for multi-label classification by attaching a classification head with five output neurons and applying a sigmoid activation function. Binary cross-entropy was used as the loss function.
  1) BERT (Bidirectional Encoder Representations from Transformers): is based on the transformer encoder and captures context from both directions of a sentence. We used the BERT-base version with 12 layers and performed full fine-tuning on all parameters. The inputs were tokenized with the original BERT tokenizer and fed through the model. [1].
  2) DeBERTa (Decoding-enhanced BERT with Disentangled Attention) by Microsoft: is based on the transformer encoder like BERT, but it improves performance by separating word content and position in the attention mechanism. It includes an enhanced decoder to predict masked words better.

We used DeBERTa-base (12 layers) and applied full fine-tuning across all layers. [2]

3) LLaMA (Large Language Model Meta AI): is a decoder-only transformer architecture developed by Meta. We used the smaller version, `Llama-3.2-1B`, which contains 1.2 billion parameters and 16 transformer layers. [7]

- **Experimentation with LLaMA:** After testing the models mentioned above, we continued with LLama for further fine-tunings and experiments ,since it was the best-performing model. Following optimization strategies are conducted to improve our results:
  - **Layer Freezing and Unfreezing:** Initially, the transformer backbone was frozen. Later We selectively unfroze the final transformer blocks (layers 13–15) to fine-tune high-level representations without overfitting.
  - **Quantization and LoRA Fine-Tuning:** The model was loaded using 8-bit quantization via `BitsAndBytesConfig`, reducing memory usage while maintaining performance. To enable fine-tuning, LoRA (Low-Rank Adaptation) adapters were applied to selected layers.
  - **Threshold Optimization:** We implemented dynamic threshold tuning for each label, instead of using fixed threshold of 0.5 for converting probabilities to binary predictions. This technique searches for optimal thresholds that can maximize the F1-score for each emotion label independently. It helps to adapt to the imbalanced labels and improves final classification performance.
- **Model Evaluation Strategy:** To evaluate the model performance, we computed micro and macro F1-scores at the end of each epoch using the validation set. Additionally, we monitored the training and evaluation??????????????????? loss across all epochs to track learning behavior and to find potential overfitting.

VI. Analysis and Interpretation

Among the three evaluated models, LLaMA achieved the best performance in both macro and micro F1 scores, shows it power in generalization for frequent and rare emotion classes.

________________________________ 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.

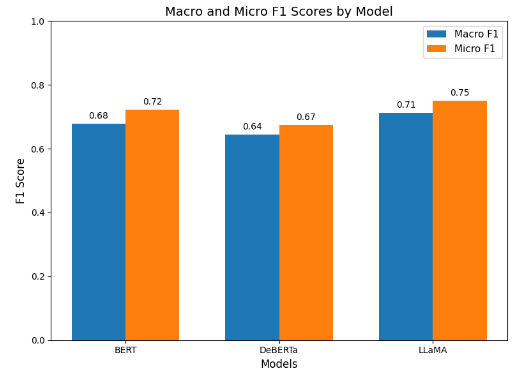

Fig. 3. Comparison of macro and micro F1 scores for BERT, DeBERTa, and LLaMA models.

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.

Model Variant	Unfrozen Layers	Macro F1	Micro F1
Llama-3.2-1B	0	0.6480	0.7117
LLaMA (3 Layers Unfrozen)	4	0.7114	0.7496
LLaMA (Quantized + LoRA)	0	0.6931	0.7421

TABLE I. Performance Comparison of LLaMA Variants on Emotion Classification

- 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 II.

Model	Epoch 1	2	3	4	5
meta-llama/Llama-3.2-1B	0.962	0.821	0.356	0.682	0.801
with 3 layer unfreezd	0.981	0.891	0.527	0.574	0.984
with 4 layer unfrezed	0.915	0.936	0.491	0.276	0.965
with 1 layer unfreezd	0.828	0.827	0.528	0.518	0.926
with 2 layer unfrrezed	0.916	0.933	0.482	0.644	0.937

TABLE II. PERFORMANCE COMPARISON OF UNFREEZING DIFFERENT LAYERS OF LLMA MODEL

VII. CONCLUSIONS AND RECOMMENDATIONS

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[4] Jiedong Lang, Zhehao Guo, and Shuyu Huang. A comprehensive study on quantization techniques for large language models, 2024.

[5] Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*, 2025.

[6] Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria, 2025. Association for Computational Linguistics.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.