Bridging the Gap in Text-Based Emotion Detection

Sarvnazsadat Roumianfar
s326344
sarvnazsadat.roumianfar@studnti.polito.it

Ali Abbasi
s323638
ali.abbasi1@studenti.polito.it

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 dataset.

1 Introduction

Emotion classification in text is a 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 [4], 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 [3], a multilingual benchmark for this task, which includes emotion-annotated texts in 28 languages. This report outlines the evolution of the project from fine-tuning traditional pretrained transformers such as BERT [1] and DeBERTa [2] to more advanced LLMs like LLaMA [5], which ultimately yielded the highest performance.

2 Method

The goal is to design a model capable of detecting multiple concurrent emotional states from short text inputs. The given task 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 the dataset and they are all considered null). Each label is independently annotated, and the model must learn to correctly identify all applicable emotions for a given sentence. 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.

  One of the properties of this subset is Class imbalance, where some emotions are overrepresented and some others are rare (see Figure 1).

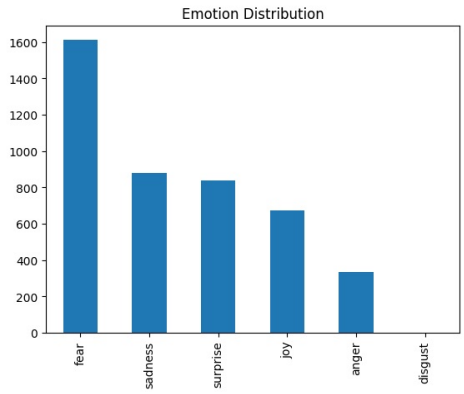

Figure 1: Emotion distribution in the English subset, illustrating class imbalance. Some emotions, such as fear, are significantly overrepresented, while others, such as anger are underrepresented.

- **Data Preprocessing:** Emotion labels were converted 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. Since Our problem is multi-label emotion detection which is a Classification problem, Encoder-Only models seems to be the best choice for this kind of tasks. So, our early attempts involved fully fine-tuning pre-trained light models such as BERT and DeBERTa. Then we move to Decoder-only models like LLaMA. Each model was adapted for multi-label classification by attaching a classification head with five output neurons and applying a sigmoid activation function.

  1. **BERT (Bidirectional Encoder Representations from Transformers):** is based on the Encoder-Only transformer 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. We have trained with batch size of 16 in epochs of 5, 15 and 20 in different runs to evaluate the performance of stability. [1].

  2. **DeBERTa (Decoding-enhanced BERT with Disentangled Attention):** is based on the transformer encoder like BERT by Microsoft, 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 with 12 layers and applied full fine-tuning on all layers. We have trained the model with a batch size of 16 across 5, 10, and 15 epochs. In one of our experiments we have applied the learning rate of 2×10^{-5} and a weight decay of 0.01. [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. We have tested batch sizes of 2, 3, 6, and 8 across 5, 8, 10, and 15 epochs. Smaller batch sizes were used to fit within the memory limitations, and larger batch sizes was because of improving training stability. We used a learning rate of 1×10^{-5} and a weight decay of 0.01. Lower learning rate was used to prevent from overshooting, which is important for fine-tuning large pre-trained models. For memory-efficient fine-tuning, we have applied 8-bit quantization with LoRA adapters. [5]

- **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 12–15), so they can learn task-specific features for emotion detection. This approach helped us fine-tune the model efficiently without overfitting.

  – Quantization and LoRA Fine-Tuning: The weights of pretrained LLaMA model are quantized to 8-bit integers (int8) instead of using full 16-bit or 32-bit floats, to reduce memory

usage. Since direct fine-tuning of quantized weights is not feasible, we applied Low-Rank Adaptation (LoRA) to the model. So, Only these adapter layers are trained and the original model stays frozen.

3 Experimental results

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. Our experiments with each model is reported below:

- **BERT** : Achieved maximum macro F1 scores around 0.66 at second epoch and after that there is sign of overfitting. (see Figure 2)

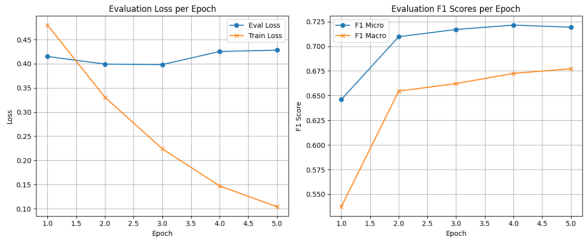

Figure 2: Bert model demonstrates steady decrease in training loss, while validation loss is increasing, resulting in overfitting. This shows that Bert model did not perform well on our task.

- **DeBERTa**: It has the similar performance like BERT. There is clear sign of overfitting after epoch 5. (See Figure 3). So changing the base model from BERT to DeBERTa does not improve the results.

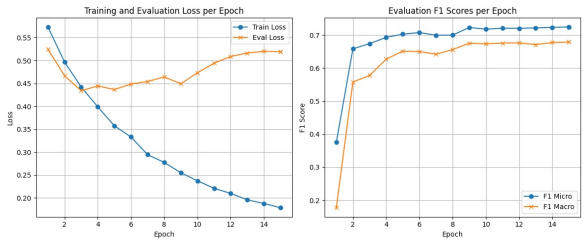

Figure 3: Training and evaluation metrics for the DeBERTa model. While training loss steadily decreases, evaluation loss begins to increase after epoch 5, indicating overfitting. Similarly, evaluation F1 scores (both micro and macro) fluctuate rather than improve, suggesting that DeBERTa did not show notably satisfying results on our task.

- **LLaMA (Full Fine-tuning)**: It has a better learning curve without signs of too much overfitting and able to achieve maximum macro F1 score of 0.67 at the 12th epoch.(see Figure 4)

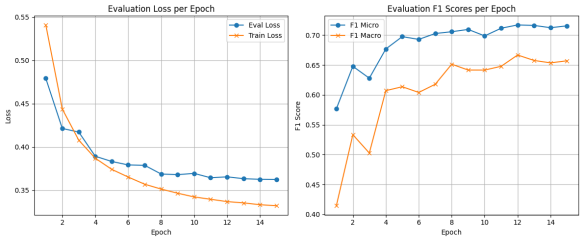

Figure 4: LLaMA demonstrates a better leaning curve than DeBERTa and BERT, indicating that it has a better learning process and better fits to our problem.

- **LLaMA (Selective Unfreezing)**: Our approach was to freeze all the layers of the LLaMA model except the last few layers. We have experimented this approach with different numbers of unfreezed layers (see Table 2). This experiment demonstrates that freezing all layers except the final 4 transformer layers dramatically improves the convergence and led to better results than others. In addition, Applying a 10x lower learning rate to these last layers prevents updating the parameters too much and improves the results. Despite being limited by our resources and only able to unfreeze a few layers, our approach is so effective that it achieves an F1 Macro accuracy of 0.71 by the fourth epoch.

4 Conclusions

In conclusion, **selective unfreezing with learning-rate separation for the unfreezed layers** is an effective technique for applying LLMs to multilabel emotion detection in low-resource scenarios. While BERT-based baselines are fast and easy to train, LLaMA outperforms them when carefully fine-tuned. Our final model achieved a Macro F1-score of 0.7114 and a Micro F1-score of 0.7496 on the English subset of the BRIGHTER dataset.

Among the three evaluated models, LLaMA achieved the best performance in both macro and micro F1 scores as shown in Table 1, demonstrates its power in generalization for frequent and rare emotion classes.

Our primary limitation was the lack of computational resources, which prevented us from training the model for a longer duration. However, our

Table 1: Training Configurations and Evaluation Results

Model	Batch Size	Epochs	LR	Weight Decay	F1-Micro	F1-Macro
BERT	16	5	2e-5	–	0.72	0.67
DeBERTa	16	15	2e-5	–	0.71	0.65
LLaMA (Quantized)	2	10	1e-5	0.01	0.74	0.69
LLaMA (4 Layer Unfrozen)	2	6	1e-5	0.01	0.74	0.71

proposed approach has showed promising results. This suggests that extending the training time and progressively unfreezing additional layers, while carefully monitoring to prevent overfitting, could further enhance performance.

Table 2: Comparison of LLaMA Variants at Epoch 8 (F1 Scores)

Model Configuration	F1-Micro	F1-Macro
Base (LLaMA-3.2-1B)	0.670	0.720
Layer 15 Unfrezeed	0.706	0.640
Layer 14,15 Unfrezeed	0.725	0.682
Layer 13 - 15 Unfrezeed	0.749	0.710
Layer 12 - 15 Unfrezeed	0.749	0.711

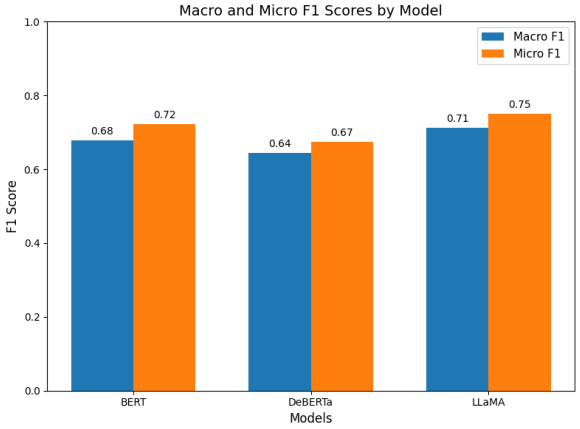

Figure 5: Comparison of macro and micro F1 scores for BERT, DeBERTa, and LLaMA with Unfreezed Layers models.

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] 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.

[4] 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.

[5] 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.