Team Unibuc - NLP at SemEval-2025 Task 11: Few-shot text-based emotion detection

Teodor-George Marchitan[1,3], **Claudiu Creanga**[2,3], **Liviu P. Dinu**[1,3]
[1] Faculty of Mathematics and Computer Science,
[2] Interdisciplinary School of Doctoral Studies,
[3] HLT Research Center,
University of Bucharest, Romania
teodor.marchitan@s.unibuc.ro, ccreanga@fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

This paper describes the approach of the Unibuc - NLP team in tackling the SemEval 2025 Workshop, Task 11: Bridging the Gap in Text-Based Emotion Detection. We mainly focused on experiments using large language models (Gemini, Qwen, DeepSeek) with either few-shot prompting or fine-tuning. With our final system, for the multi-label emotion detection track (track A), we got an F1-macro of 0.7546 (26/96 teams) for the English subset, 0.1727 (35/36 teams) for the Portuguese (Mozambican) subset and 0.325 (**1**/31 teams) for the Emakhuwa subset.

1 Introduction

Task 11 from the International Workshop on Semantic Evaluation (Muhammad et al., 2025b) focuses on identifying emotions that most people would think the speaker might feel given a short piece of text. With all the advances in the AI world, automatic emotion detection plays such an important role: it can lead to more empathetic and context-aware interactions between humans and chatbots / AI assistants, improving user experience and satisfaction; it can help monitoring mental health conditions and provide timely interventions such that more people seek for help before is too late; it can help business to understand the customers sentiments regarding different products or services and can improve their strategies.

The best system developed for Task 11 track A is based on Gemini Flash (et. al., 2024) model using different techniques of few-shot prompting.

We made our models publicly available in a GitHub Repository.

2 Background

The competition is multilingual and it had 3 tracks:

A. Multi-label Emotion Detection (28 supported languages)

B. Emotion Intensity (12 supported languages)

C. Cross-lingual Emotion Detection (28 supported languages)

We participated to track A with our system for 3 languages: Emakhuwa 0.325 F1-macro (position **1**/31 teams), English 0.7546 F1-macro (position 26/96 teams) and Portuguese (Mozambican) 0.1727 F1-macro (position 35/36 teams).

2.1 Dataset

The BRIGHTER dataset (Muhammad et al., 2025a) was collected from 4 different data sources (social media posts, personal narratives, literary texts, and news data) in 28 different languages:

Language	Train	Dev	Test
ENG	2,768	116	2,767
PTMZ	1,546	257	776
VMW	1,551	258	777

Table 1: Train/Dev/Test splits for languages tackled with our system for track A

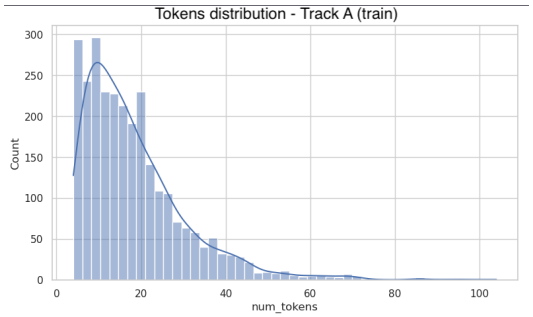

Figure 1: Track A: Distribution of token length for the training dataset in english

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 468–475

2.2 Previous Work

Emotion detection in text has been a subject of study for a considerable time, with early approaches relying on lexicon-based methods. These methods leverage pre-compiled lists of words associated with different emotions. For instance, the presence of words like "happy," "joyful," or "excited" might indicate a positive emotion, while words like "sad," "depressed," or "miserable" could suggest sadness. Later, traditional machine learning classifiers such as Support Vector Machines (SVMs), Naive Bayes, and Maximum Entropy models became popular for emotion detection. These models typically rely on handcrafted features extracted from text, including bag-of-words, TF-IDF, n-grams, and sentiment lexicons.

More recently, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), proved effective in capturing sequential information in text, enabling models to better understand context and dependencies within sentences and documents (Poria et al., 2017). Convolutional Neural Networks (CNNs) were also employed to extract local features and patterns indicative of emotion (Kim, 2014).

Transformer-based models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and their variants, have achieved state-of-the-art performance in various NLP tasks, including emotion detection. These models, pre-trained on massive amounts of text data, capture rich contextual representations of words and sentences, leading to significant improvements in accuracy and robustness. Fine-tuning these pre-trained models on emotion-annotated datasets became a common practice for achieving high performance in emotion detection tasks. The SemEval challenges have consistently played a significant role in driving research in emotion detection. SemEval 2024 contained Task 10 which tackled emotion discovery and reasoning flip in conversation (EDiReF). The majority of participants used LLMs and achieved best results (Creanga and Dinu, 2024).

3 System overview

In this paper, we focused our research on three approaches: **fine-tuning BERT** based models (3.1), and two techniques for LLMs: **Few-Shot Prompting** (3.2) and **Fine-tuning** (3.3).

3.1 Fine-tuning BERT based models

We experimented with several prominent transformer-based models: DeBERTa v3 Large (He et al., 2021), mBERT (Multilingual BERT) (Devlin et al., 2018), and XLM-RoBERTa-large (Conneau et al., 2019). These models represent advancements in pre-trained language model architectures and have demonstrated strong performance in cross-lingual understanding tasks. We utilized these pre-trained models as feature extractors and appended a simple classification head on top. We extract sentence-level feature representations by specifically using the [CLS] token output from the transformer's final hidden layer. To prevent overfitting, we apply a dropout rate of 0.3 to these features. These extracted features are then fed into a fully connected network, designed for classification. This network comprises two core blocks. Each block progressively reduces the feature dimensionality, starting with 512 neurons and narrowing down to 128. The final layer of the fully connected network has 6 output neurons, corresponding to the 6 emotion categories we are predicting. We apply a higher dropout rate of 0.5 within the fully connected network, again for regularization. The model outputs raw prediction scores without any activation function.

We trained this model over 3 epochs in a two-stage approach, as recommended in (Marchitan et al., 2024). Initially, for the first 2 epochs, we froze the transformer layers, only training the weights of our fully connected classification network. This allows the classification layers to adapt to the pre-trained transformer features without disrupting the transformer's learned representations. In the final 1 epoch, we unfroze the last transformer layer and fine-tuned it along with the fully connected network. This enables the model to slightly adjust the transformer's understanding of the input text to be even more relevant for our specific emotion detection task. We used a batch size of 32 during training and employed the AdamW optimizer, known for its effectiveness in training transformers. For the initial frozen-transformer training phase, we used a learning rate of 3e-4, which was reduced to 2e-4 during fine-tuning to prevent destabilizing the already partially trained model. A linear learning rate scheduler with a 50-step warm-up was used to gradually increase the learning rate at the beginning of training, promoting stable convergence. We used cross-entropy

loss as our objective function.

While our BERT-based models, provided a good baseline, their performance did not reach the levels achieved by LLMs like Gemini Flash (Table 2). However, it's important to note that these BERT models offer a significant advantage in terms of computational resources. They are considerably smaller in model size and demonstrate substantially faster inference speeds compared to LLMs.

Lang	Model	F1 Macro	F1 Micro
ENG	mBERT	0.54	0.56
ENG	DeBERTa	0.70	0.70
ENG	XLM-RoBERTa	0.70	0.71

Table 2: Results of BERT based models for validation set.

3.2 Few-Shot Prompting

We found that Few-Shot prompting beats Zero-Shot by around 7% in performance. Prompts can be inspected in Appendix A. For example Qwen2.5-14B Zero-Shot CoT obtained an F1 Macro of 0.63, while with Few-Shot 0.70. This was observed across all models. Another insight was that increasing the number of examples we gave to the model will result in an increase in performance. We tried giving 6 examples, one that included each emotion and progresively giving it more examples: 100, 300 and 600. Best results were when we gave the most examples (600). A third insight is that we found out that complex prompting techniques like CoT and ToT actually make the models perform worse. We believe this is because the task is a perceived emotion task, where the annotations are not necesarily the ground truth, but what the labelers considered to be true. Following the thinking process of the models, in the examples where they disagreed with the ground truth, we were actually conviced that the models were right in most of the cases.

Our experiments (Table 3) revealed a consistent performance advantage for few-shot prompting over zero-shot prompting, with an approximate improvement of 7% in F1-macro. For example, the Qwen2.5-14B model (Qwen et al., 2025) achieved an F1-macro score of 0.63 using zero-shot prompting, while few-shot prompting boosted performance to 0.70. This trend was observed across all models evaluated. Furthermore, we investigated the impact of increasing the number of examples provided within the few-shot prompts. By progressively increasing the example count (from an initial set of 6 examples, one for each emotion, to 100, 300, and finally 600 examples), we observed a positive correlation between the number of examples and performance. The highest F1-macro scores were consistently obtained when utilizing the largest example set of 600.

Counterintuitively, we found that employing more complex prompting techniques such as Chain-of-Thought (CoT) (Wei et al., 2022) and Tree-of-Thought (ToT) (Long, 2023) did not yield the expected performance gains. In fact, these methods tended to slightly degrade performance in our specific task. We hypothesize that this unexpected outcome stems from the nature of the "perceived emotion" task itself. In this task, annotations do not necessarily represent an objective "ground truth" emotion, but rather reflect human annotators' subjective interpretations of the speaker's likely feelings. Consequently, while CoT and ToT are designed to encourage models to mimic human-like reasoning processes, in this context, forcing the model to explicitly articulate a detailed thought process may not align with the inherently subjective and potentially less consciously reasoned nature of perceived emotion annotation. Indeed, in instances where model predictions diverged from the assigned labels, our qualitative analysis suggested that the models' reasoning, often aligned with alternative, yet plausible, emotional interpretations of the text.

Model	Type	Examples	F1 Macro
Qwen2.5	Zero-Shot	0	0.63
Qwen2.5	Z-S CoT	0	0.63
Gemini	F-S ToT	6	0.63
Gemini	F-S ToT	20	0.64
Gemini	Few-Shot	6	0.69
Qwen2.5	Few-Shot	6	0.70
Gemini	Few-Shot	500	0.76
Gemini	Few-Shot	600	0.77

Table 3: Results of LLM models for validation set. The exemples were given from the training set. We used Qwen2.5-14B and Gemini 2.0 Flash Exp. Z-S means Zero-Shot, F-S means Few-Shot.

3.3 Fine-tuning

We experimented fine-tuning with different large language models (DeepSeek (et. al., 2025), Mis-

tral 7B (Jiang et al., 2023) and Qwen2.5 (Qwen et al., 2025)) as the backbone of our system architecture, on top of which we added a multi-label classification layer in order to classify the emotions present in the given text. We set the maximum number of tokens to 256 and we truncated the longer text by keeping the first part of the text, as suggested in (Marchitan et al., 2024). We then fine-tuned the quanitized model using Low-Rank Adaptation (LoRA) (Hu et al., 2022) with following hyperparameters: $r = 4$ ($r = 2$ for Mistral 7B), $lora\_alpha = 8$ and $lora\_dropout = 0.05$ ($lora\_dropout = 0.1$ for Qwen2.5-1.5B). The fine-tuning was done using the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 0.01, weight ratio of 1% for Mistral 7B and Qwen2.5-1.5B, but 2% for other 2 models. We set the learning rate to $2e-5$ when fine-tuning DeepSeek and Qwen2.5-1.5B, $5e-5$ for Mistral 7B and $2e-4$ for Qwen2.5-0.5B. Cross entropy loss was used during training to measure the performance. We used different batch sizes (4 for Mistral 7B; 8 for DeepSeek and Qwen2.5-0.5B; 32 for Qwen2.5-1.5B) based on the total number of trainable parameters in order to be able to fit on the available hardware. We set the maximum number of epochs to 15, but the actual number of training epochs varies as we implemented the early stopping strategy and the final model is the one with the smallest loss on the validation set.

LLM Backbone	Epochs	Train Loss	Dev Loss
DeepSeek R1 Distill Llama 8B	5	0.2495	0.3181
Mistral 7B	3	0.2515	0.3455
Qwen2.5-0.5B	3	0.3941	0.3614
Qwen2.5-1.5B	8	0.3714	0.3571

Table 4: Train and validation losses for each model alongside the number of epochs trained.

4 Results

We participated in Track A and tested our proposed system on 3 languages: English, Portuguese (Mozambican) and Emakhuwa. We managed to be over the baseline on 2 out of the 3 languages and secured the first position for one of them, as shown in Table 6. This reflects our model's ability to adapt with few-shot in-context learning especially on languages with limited availability of NLP resources (such as Emakhuwa).

Language Code	F1 Macro	Place
VMW	0.3250	**1** / 31
ENG	0.7546	26 / 96
PTMZ	0.1727	35 / 36

Table 6: Team Unibuc - NLP results on Track A.

4.1 Error Analysis

Examining the confusion matrices (Figure 2, Figure 3, Figure 4, Figure 5, Figure 6), we observe distinct performance profiles across emotions. For fear, while the model achieves its highest True Positive rate (TP) at 88%, indicating strong detection of actual fear, it simultaneously exhibits its lowest True Negative rate (TN) at 70%. This combination suggests a tendency to over-predict fear, potentially misclassifying other emotions as fear. Conversely, for joy, the model shows the weakest performance in detecting the emotion when it is present, achieving the lowest TP of 67% and consequently, the highest False Negative rate (FN) of 32%, indicating a higher likelihood of missing instances of joy. Furthermore, the highest False Positive rate (FP) of 30% is also associated with fear, reinforcing the observation of potential over-prediction for this emotion. It seems like with both approaches we have experimented with LLMs (few-shot prompting and fine-tuning) the models tend to over-predict fear while under-predicting joy (Figure 7).

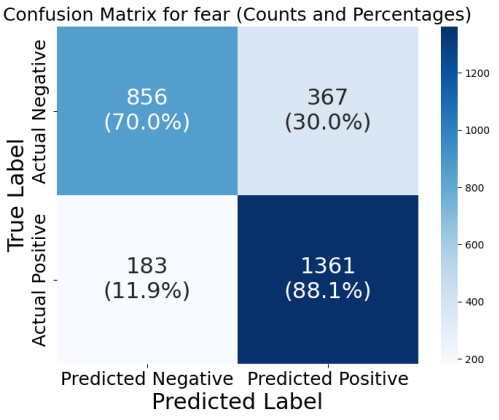

Figure 2: Confusion Matrix for Fear (Counts and Percentages)

LLM Backbone	Validation F1 Micro	Validation F1 Macro	Test F1 Micro	Test F1 Macro
DeepSeek R1 Distill Llama 8B	0.7723	0.7602	0.7744	0.7441
Mistral 7B	0.7696	0.7305	0.7699	0.7268
Qwen2.5-0.5B	0.7496	0.6975	0.7224	0.6840
Qwen2.5-1.5B	0.7340	0.6906	0.7319	0.6826

Table 5: F1 Micro and F1 Macro results on both train and validation datasets.

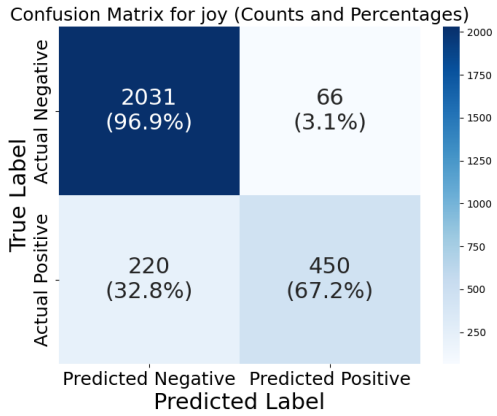

Figure 3: Confusion Matrix for Joy (Counts and Percentages)

5 Conclusions and Future Work

Our investigation explored three primary approaches: fine-tuning BERT-based models and leveraging LLMs through both few-shot prompting and fine-tuning techniques. While fine-tuned BERT models provided a solid performance baseline, they were ultimately surpassed by methods employing LLMs, particularly Gemini Flash with few-shot prompting, which formed the basis of our best-performing system. Interestingly, complex prompting strategies like Chain-of-Thought and Tree-of-Thought did not yield improvements, but degradations, possibly due to the subjective nature of the perceived emotion task.

Building upon the insights gained from this study, several avenues for future research emerge. Firstly, further investigation is warranted to understand the performance disparity observed across languages, particularly the lower F1-macro score for the Portuguese subset. Detailed dataset analysis and error analysis could reveal language-specific nuances or data biases impacting model performance. Secondly, exploring the new "thinking models" like "Gemini 2.0 Thinking" and "Thinking Claude" should be tested, as they show promises in other similar tasks.

Acknowledgements

Research partially supported by the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, SiRoLa project, PN-IV-P1-PCE-2023-1701, and InstRead project PN-IV-P2-2.1-TE-2023-2007, within PNCDI IV, Romania.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Claudiu Creanga and Liviu P. Dinu. 2024. ISDS-NLP at SemEval-2024 task 10: Transformer based neural networks for emotion recognition in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 649–654, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

DeepSeek-AI et. al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Gemini Team et. al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

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.

Jieyi Long. 2023. Large language model guided tree-of-thought.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Teodor-george Marchitan, Claudiu Creanga, and Liviu P. Dinu. 2024. Team Unibuc - NLP at SemEval-2024 task 8: Transformer and hybrid deep learning based models for machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 403–411, Mexico City, Mexico. Association for Computational Linguistics.

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. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages.

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. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

A Example of prompts

A.1 Zero shot

Analyze the following sentence and identify all emotions that are present. Select from this list: "Anger", "Fear", "Joy", "Sadness", "Surprise". If multiple emotions are present, list them separated by commas. If no emotions from the list are present, respond with "None".

Sentence: [Sentence]

Emotions:

A.2 Zero shot with CoT

Analyze the following sentence and identify all emotions that are present. Select from this list: "Anger", "Fear", "Joy", "Sadness", "Surprise". If multiple emotions are present, list them separated by commas. If no emotions from the list are present, respond with "None". Let's break down the emotional content step by step.

Sentence: [Sentence]

Reasoning: Consider the specific words used, the context of the sentence, and any implied feelings.

Emotions:

A.3 Few shot with CoT

Analyze the following sentence and identify all emotions that are present.

Examples: 1. Sentence: "But not very happy." Emotions: Joy, Sadness 2. Sentence: "They were dancing to Bolero" Emotions: Joy, Sadness 3. Sentence: Yes, the Oklahoma city bombing." Emotions: Anger,Fear,Sadness,Surprise 4. Sentence: "5 year old me was scarred for life." Emotions: Fear, Sadness 5. Sentence: "How stupid of him." Emotions: Anger 6. Sentence: "I turned around so I could see my back." Emotions: Surprise

Given the following sentence, select from this list: "Anger", "Fear", "Joy", "Sadness", "Surprise". If multiple emotions are present, list them separated by commas. Let's break down the emotional content step by step.

Sentence: [Sentence]

Reasoning: Consider the specific words used, the context of the sentence, and any implied feelings. Output only the emotions that are present. No other words.

Emotions:

A.4 Few Shot with ToT

Analyze the following sentence and identify all emotions that are present. Given the following sentence, select from this list: "Anger", "Fear", "Joy", "Sadness", "Surprise".

Examples: [600 examples]

Given the following sentence, select from this list: "Anger", "Fear", "Joy", "Sadness", "Surprise". If multiple emotions are present, list them separated by commas. Let's break down the emotional content step by step using a Tree of Thoughts approach.

Sentence: [Sentence]

Reasoning:

Thought 1: Initial Impression: What is the first emotion that comes to mind upon reading the sentence? Briefly explain why.

Thought 2: Word-Level Analysis: Are there any specific words or phrases that strongly suggest an emotion? If so, which words and which emotions?

Thought 3: Contextual Considerations: Does the context of the sentence provide any additional clues about the emotional state? Consider the situation being described.

Thought 4: Alternative Interpretations: Are there any other possible interpretations of the sentence that might suggest different emotions? Explore these possibilities.

Thought 5: Synthesis: Based on the previous thoughts, which emotions are most likely present in the sentence? Justify your final selection.

Final Emotions: Output only the emotions that are present, separated by commas. No other words.

Emotions:

B Error analysis

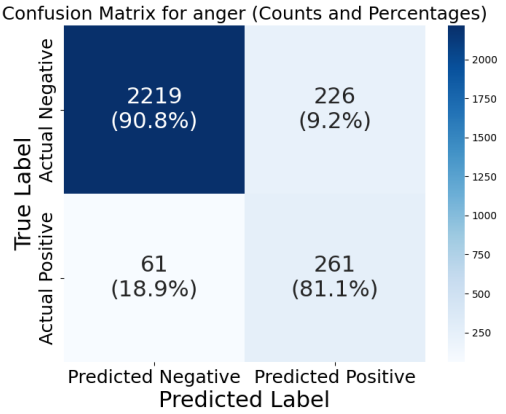

Figure 4: Confusion Matrix for Anger (Counts and Percentages)

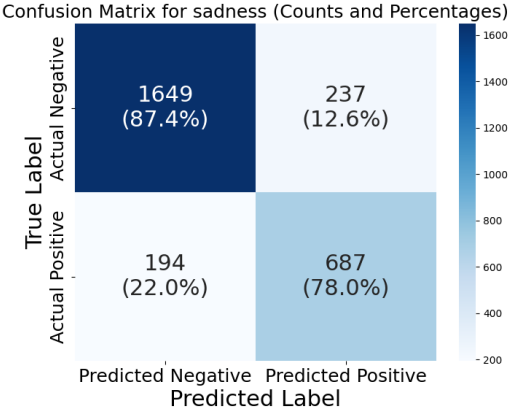

Figure 5: Confusion Matrix for Sadness (Counts and Percentages)

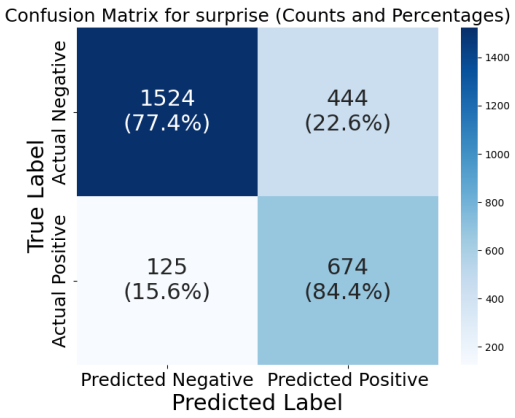

Figure 6: Confusion Matrix for Surprise (Counts and Percentages)

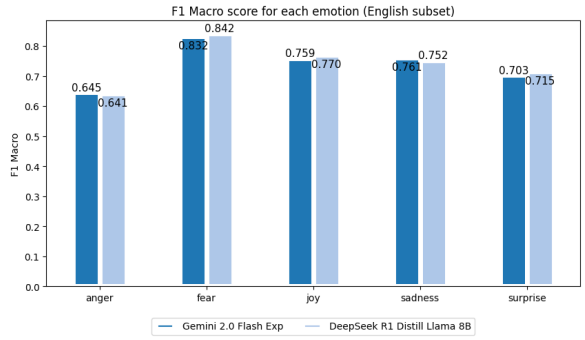

Figure 7: Comparison of F1 macro score on each emotion between best model using few-shot prompting (Gemini 2.0 Flash) and fine-tuning (DeepSeek R1 Distill Llama 8B).