Emotion Train at SemEval-2025 Task 11: Comparing Generative and Discriminative Models in Emotion Recognition

Anastasiia Demidova[λ], **Injy Hamed**[λ], **Teresa Lynn**[λ], **Thamar Solorio**[λ,ξ]
[λ]MBZUAI, Masdar City, SE45 05, Abu Dhabi, UAE
[ξ]University of Houston, Houston, TX 77204-3010, USA
{anastasiia.demidova, injy.hamed, teresa.lynn, thamar.solorio}@mbzuai.ac.ae

Abstract

The emotion recognition task has become increasingly popular as it has a wide range of applications in many fields, such as mental health, product management, and population mood state monitoring. SemEval 2025 Task 11 Track A framed the emotion recognition problem as a multi-label classification task. This paper presents our proposed system submissions in the following languages: English, Algerian and Moroccan Arabic, Brazilian and Mozambican Portuguese, German, Spanish, Nigerian-Pidgin, Russian, and Swedish. Here, we compare the emotion-detecting abilities of generative and discriminative pre-trained language models, exploring multiple approaches, including curriculum learning, in-context learning, and instruction and few-shot fine-tuning. We also propose an extended architecture method with a feature fusion technique enriched with emotion scores and a self-attention mechanism. We find that BERT-based models fine-tuned on data of a corresponding language achieve the best results across multiple languages for multi-label text-based emotion classification, outperforming both baseline and generative models.

1 Introduction

The task of emotion recognition involves identifying emotions in text or speech. Track A of SemEval 2025 Task 11 (Muhammad et al., 2025b) focuses on multi-label emotion recognition in social media texts across several languages. Following the definition of the universal emotions introduced by Ekman (1992), the task involves classifying the texts for the following six basic emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise. In this paper, we present our work submitted to the shared task for the following languages: English, Algerian and Moroccan Arabic, Brazilian and Mozambican Portuguese, German, Spanish, Nigerian-Pidgin, Russian, and Swedish.

Large language models (LLMs) have achieved remarkable results on a wide range of applications (Chang et al., 2024). As these models become increasingly integrated into real-world settings covering diverse domains, LLMs are expected to exhibit human-like behaviour for proficient social interactions. This served as a motivation to include LLMs in our investigated approaches, comparing their performance to discriminative pre-trained language models (PLMs). Accordingly, our approaches fall under two main tasks, *Classification* and *Generation*, as we work with both discriminative and generative PLMs. We explore a wide range of techniques, including zero- and few-shot prompting, as well as fine-tuning and few-shot fine-tuning. We also propose a novel approach to extending the BERT architecture with feature fusion and self-attention, incorporating token-level emotion scores statistically calculated on the train set.

In our conducted experiments, non-causal models demonstrated superiority over causal ones. We also observed that imbalanced data has a high impact on a model's performance, notably biasing outcomes toward the detection of 'Fear' in the English setup. Our code is available online.[1]

2 Related Work

Emotion Recognition. Strapparava and Mihalcea (2007) and Mohammad et al. (2018) addressed emotion recognition in the SemEval challenges, tackling tasks such as affective text exploration, bridging emotional and sentiment aspects, and inferring speakers' emotions from tweets. More recently, Zhang et al. (2023) examined an architecture with discourse- and speaker-aware modules within graph attention networks, which outperformed the state-of-the-art (SOTA) in the task of Emotion Recognition in Conversations.

Nag et al. (2023) explored several deep learning

[1] https://github.com/profii/semeval25_task11

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1004–1014

techniques to address different emotional intelligence (EI) tasks, including emotion recognition. Zhao et al. (2024) tackled the issue of catastrophic forgetting in LLMs, which was previously reported by Luo et al. (2023) when integrating EI.

PLMs have proven to be highly effective on various NLP benchmarks (Sun et al., 2019; Devlin et al., 2019; OpenAI, 2023; Nikishina et al., 2023; Chowdhery et al., 2023; Demidova et al., 2024; etc.). With respect to the current task, we considered the following approaches:

Fine-tuning. Recent advances have been made in fine-tuning approaches by adapting PLMs with minimal parameter updates. For example, PEFT (Ding et al., 2023) and LoRA (Hu et al., 2022) techniques significantly reduce the computational requirements while maintaining high performance.

In-Context Learning. Zero- and few-shot ICL offer the advantage of not modifying the model parameters. Dong et al. (2024)'s survey provides a taxonomy of ICL that demonstrates various ways to apply pre-trained language models in NLP tasks. Brown et al. (2020) highlight the effectiveness of few-shot ICL reaching SOTA performance.

Few-shot Fine-tuning. Mosbach et al. (2023) presented a method of few-shot fine-tuning that is between ICL and full fine-tuning. Their approach involves using a small number of labelled examples in the input during the fine-tuning stage (resembling few-shot learning).

Feature Fusion with Self-Attention. The feature fusion (FF) method implies a combination of multiple feature representations, such as embeddings. Recent works (Yang et al., 2020, 2024) showed implementations of FF under self-attention that enhanced model performance in Named Entity Recognition. Santoso et al. (2021) explored the combination of self-attention and word-level features that improve Speech Emotion Recognition.

3 Methodology

In this section, we describe the system overview by examining both groups of approaches for classification and generation tasks in detail.

3.1 Data

The organizers of the SemEval 2025 Task 11 (Muhammad et al., 2025a) provided 28 datasets across different languages, taking as resources news, social media, annotated speeches, translations from literature, and examples written by natives and augmented with machine-generated content. To simplify the annotation process, the authors chose the following six labels: Anger, Disgust, Fear, Joy, Sadness, and Surprise. They did not include Disgust in the English dataset due to the insufficient number of class elements. In our work, we only use the training dataset from Muhammad et al. (2025a) to train our models, while the development set is used for evaluation. Table 4 in Appendix A.1 provides an analysis of the task's datasets. As further training data, we also considered GoEmotions (Demszky et al., 2020) and MELD (Poria et al., 2019), but early experiments showed that they did not offer any improvement, which we believe is because their annotations represent speakers' emotions instead of perceived ones.

As one utterance can evoke several emotions, we analyzed all emotion combinations occurring in the data. We provide further analysis for English in Appendix A.1.2, showing emotions co-occurrence.

Preprocessing. The informal nature of the social media domain presents noisy content such as hashtags, mentions, emojis, elongated words, informal abbreviations, and various punctuation styles. While these elements can help in the expression of emotions, they can also complicate the tokenization process. We believe that preprocessing can help improve the consistency of textual representations for emotion classification. To clean our data, we perform the following preprocessing steps: standardizing similar emojis to a set of basic Ekman emotions (Anger: ':@', Fear: 'D:', Joy: ':)', Sadness: ':(', Surprise: ':o', Neutral: ':|'), as well as removing elements such as URLs, user mentions and hashtags[2]. We evaluated the effectiveness of these steps across a subset of the languages, revealing a benefit to English only.

3.2 Models

We explore the use of non-causal and causal PLMs, thus organizing our approaches into two main categories: Classification and Generative.

3.2.1 Baselines

In the monolingual setup, organizers of Muhammad et al. (2025a) experimented with chain-of-thought prompting on various LLMs (Qwen2.5-

[2]The preprocessing script is provided in the GitHub repository: https://github.com/profii/semeval25_task11

72B, Dolly-v2-12B, Llama-3.3-70B, Mixtral-8x7B, and DeepSeek-R1-70B) and fine-tuning on multilingual language models (LaBSE, RemBERT, XLM-R, mBERT, and mDeBERTa).

3.2.2 Classification Models

We utilize RoBERTa-large for the English setup (Liu et al., 2019) and XLM-RoBERTa-large for the other languages (Conneau et al., 2020). These models are optimized for contextual relationship understanding and include an attention mechanism, making them suitable for text classification tasks.

Fine-tuning. We fine-tune non-causal models, where the hyper-parameter values are specified in Appendix A.2.1.

Curriculum Learning. Curriculum Learning is a fine-tuning approach designed to improve model performance by starting with easier examples and progressively introducing more complex ones. We apply this strategy by beginning with neutral utterances, followed by examples containing only one emotion, and gradually increasing the complexity to sentences having multiple emotions.

Feature Fusion with Self-Attention. Motivated by the work of Santoso et al. (2021), we integrate emotional features to enhance the model's ability to capture nuanced emotional contexts and assign dynamic weights. This extension recognizes that different tokens hold various levels of emotional relevance. We apply two transformations to the model architecture during fine-tuning: additional two layers (self-attention and linear classifier) and the token-level feature fusion with emotion scores, as shown in Figure 1. The self-attention layer takes as input token-level emotion-weighted embeddings that are concatenated to the embeddings produced by the previous layer. The emotion-weighted embedding is equal to the number of emotion classes, where each element contains an emotion score, representing a probability of that emotion being associated with that input token. In order to calculate the emotion scores, we first tokenize the sentences in the training data using the relevant non-causal model for each language. For each token, the emotion score is based on the number of occurrences of that token (e.g. 'tears') in sentences with a certain emotion label (e.g. Fear, Sadness), divided by the total occurrences of this token across all emotions.[3]

[3]Our emotion score is calculated on the training data.

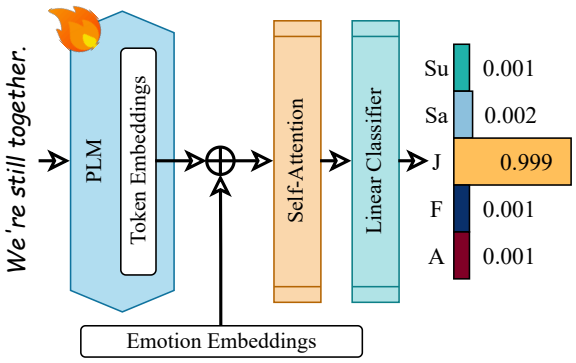


Figure 1: Scheme of the Feature Fusion with Self-Attention (FFSA) approach, where $\oplus$ denotes the concatenation operator and *A, F, J, Sa*, and *Su* represent Anger, Fear, Joy, Sadness and Surprise, respectively.

Figure 2 provides an example of an emotion score mapping.

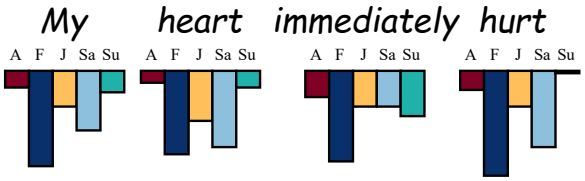


Figure 2: Example of Emotion Score mapping, where *A, F, J, Sa*, and *Su* represent Anger, Fear, Joy, Sadness and Surprise, respectively.

3.2.3 Generative Models

We utilize the following LLMs: Meta-Llama-3-8B-Instruct (Llama-3) (Grattafiori et al., 2024), GPT-4o-mini (OpenAI, 2024), and Qwen2.5-32B-Instruct (Qwen2.5) (Qwen, 2025). Our choice of causal models is based on preliminary experiments that we conducted, where these LLMs outperformed others.[4] The superiority of these models has also been confirmed by Wang et al. (2023), where they demonstrated both high EI and strong performance on general tasks.

We explore the use of in-context learning and fine-tuning. For both approaches, the inference part consists of generating output and deducing the emotion. Table 7 in Appendix A.2.2 demonstrates the hyper-parameters for the inference mode, where we opted for less creative and more consistent responses. As generative models can provide

[4]We also experimented with Mistral-7B-v0.1, Qwen2.5-14B-Instruct, Phi-3-medium-4k-instruct, Meta-Llama-3-8B, deepseek-ai/deepseek-llm-7b-chat, Gemma-2-9b, Llama-3.1-8B-Instruct, and Vicuna-7b-v1.5.

Language	Model	BERT-F1	XLM-F1	Llama-F1
English	RoBERTa-large	**78.9**	76.5	73.0
Algerian Arabic	MarBERT	**58.5**	46.8	39.3
Moroccan Arabic	bert-base-arabic-camelbert-msa	**56.3**	51.9	34.4
Brazilian Portuguese	bert-base-portuguese-cased-large	53.1	**53.4**	37.9
Mozambican Portuguese	bert-base-portuguese-cased-large	**54.8**	52.0	32.0
German	gbert-large	68.2	**69.9**	42.2
Nigerian-Pidgin	-	-	**58.0**	34.5
Russian	RuBERT-large	83.9	**85.4**	49.9
Spanish	RoBERTa-BNE-base	77.0	**77.2**	54.5
Swedish	bert-base-swedish-cased	**50.8**	23.5	37.3

Table 1: Best results on the **development set**, showing the F1-Macro scores for BERT-based language-specific models (*BERT-F1*), XLM-R (*XLM-F1*), and Llama-3-8B-Instruct (*Llama-F1*). Best models are bolded.

final responses outside of the given set of emotions, we apply a post-processing step to map the model output to the six Ekman emotions, using the GoEmotions mapping (Demszky et al., 2020) provided in Table 8 in Appendix A.3.

In-Context Learning. In the context of limited GPU memory, 8-bit quantization with bitsandbytes[5] allowed us to experiment with multiple prompt templates. In Appendix A.4, we demonstrate the two most effective prompts for English that we subsequently use throughout all our ICL experiments, as well as the examples we used for few-shot learning.[6]

Instruction Fine-tuning. We perform instruction fine-tuning on Llama-3. Due to computational limitations, experiments are conducted with 4-bit quantization and the LoRA adapter. Fine-tuning hyper-parameters are also specified in Appendix A.2.2. To achieve higher performance, we additionally implemented few-shot fine-tuning.

3.3 Evaluation Metrics

We report macro-averaged F1 score, which is the main metric used for evaluation by the shared task organizers (Muhammad et al., 2025a). F1-Macro is defined as the (unweighted) average of F1 scores calculated separately for each label.

4 Results

For discriminative models, we use BERT-based language-specific models as well as XLM-R, covering fine-tuning, FFSA, and CL across all languages. For generative models, we conduct preliminary experiments on the English language, where the best setup was found to be using Llama-3 along with prompt#1 (see Appendix A.4). In Table 2, we present the best results across different generative models for English. Due to resources constraints for experimenting with other languages, we apply this best-performing setting across all languages. Table 1 presents the best results for language-specific BERT models, XLM-R, and Llama-3 on the development set for each language. In Appendix A.5, we elaborate on experimental results on the English setup.

Model	Prompt #	F1-Macro
Llama-3	1	**73.0**
GPT-4o-mini	2	69.8
Qwen2.5	2	69.1

Table 2: Best results on Prompting for the English development set, where *Llama-3* is Meta-Llama-3-8B-Instruct, *Qwen2.5* - Qwen/Qwen2.5-32B-Instruct.

In Table 3, we present the results on the test set. We demonstrate a comparison between our results, the best scores from Muhammad et al. (2025a), and the highest F1-Macro in competition across different languages.

Our results show that discriminative models outperformed Llama-3 across all languages. Among discriminative models, we find that XLM-R is more consistent across multiple languages except for Arabic and Swedish, considering the difference between XLM-R and language-specific BERTs.

5 Discussion

In order to further understand our experiment results on the English development set, we analysed confusion matrices of models with fine-tuning and Feature Fusion with Self-Attention (FFSA) approaches (see Figures 7a and 7b in Appendix A.6.1). We observed that both models are overfitting by choosing 'Fear' in most of the con-

[5] https://huggingface.co/docs/bitsandbytes/

[6] For all prompt-related experiments, the model was prompted in English with a language-specific example.

Language	Model	Approach	Baseline*	F1-Macro†	BEST‡	Δ
Russian	XLM-R-large	Fine-tuning	83.8	**88.7**	90.9	2.2
Mozambican Portuguese	portuguese-large	Curriculum learning	45.9	**50.7**	54.8	4.1
Moroccan Arabic	camelbert-msa	Fine-tuning	52.8	**57.8**	62.9	5.1
English	RoBERTa-large	FFSA	70.8	**76.2**	82.3	6.1
Spanish	XLM-R-large	Fine-tuning	77.4	**77.8**	84.9	7.1
Swedish	swedish	Curriculum learning	52.0	**55.3**	62.6	7.3
Nigerian-Pidgin	XLM-R-large	Fine-tuning	55.5	**60.0**	67.4	7.4
German	gbert-large	Fine-tuning	64.2	**66.0**	74.0	8.0
Algerian Arabic	MarBERT	Fine-tuning	55.8	**57.9**	66.9	9.0
Brazilian Portuguese	portuguese-large	FFSA	51.6	**54.7**	68.3	13.6

Table 3: We report the best results we achieve on the **test set** across languages. We present the best result of baselines from Muhammad et al. (2025a) ($\star$), our results ($\dagger$), the highest F1-Macro in competition ($\ddagger$), and the difference between the best in competition and our score (Δ). We use the following abbreviations: *FFSA* for feature fusion with self-attention approach, *camelbert-msa* for bert-base-arabic-camelbert-msa, *portuguese-large* for bert-base-portuguese-cased-large, and *swedish* for bert-base-swedish-cased.

fusion cases, possibly due to imbalanced data (see also Figure 4 in Appendix A.1.3). Regarding FFSA, the additional self-attention layer appears to improve the distinction between Anger, Fear and Joy. Additionally, the dataset samples demonstrate that the fine-tuned RoBERTa model does not effectively distinguish Sadness and Surprise emotions from others, interpreting them as Anger or Fear and Fear or Joy, respectively, due to ambiguous cases.

Moreover, a comparison of these two approaches on F1, Recall, and Precision scores with Statistical Significance (Berg-Kirkpatrick et al., 2012) shows a 0.84 P-value, which means that the difference in the performance of these two models is not statistically significant. However, the FFSA technique does not require much additional high computational power, allowing this method to be applied efficiently in the English setup.

We believe multiple factors could be affecting the performance of models across languages, including data imbalance, sentence length and dataset size. Mozambican Portuguese and Algerian Arabic demonstrate the lowest results, likely due to being low-resource languages with relatively small datasets. In contrast, for Nigerian-Pidgin, despite typically being a low-resource language, Nigerian-Pidgin XLM-R performs relatively well. We believe this could be due to being well-resourced in this set-up (see Table 4), as well as the prevalence of English in the language. In terms of sentence length, Brazilian Portuguese and Swedish contain longer sentences (on average and at their maximum lengths), complicating input processing. As for German, its linguistic similarities to English suggest strong model performance. However, we believe models might struggle with longer dependencies and complex sentence structures due to relatively long sentences.

Regarding error analysis for RoBERTa, Expected Calibration Error (ECE) of the approaches of both fine-tuning and fine-tuning with Curriculum Learning and Data Preprocessing have 8.7% and 8.5% ECE, respectively. This indicates that these models are well-calibrated but still have miscalibration. As for the Feature Fusion with Self-Attention approach, the model is on a threshold with 10.3% ECE, meaning the model's confidence levels might be overconfident or underconfident, compared to actual outcomes. Comparing selected predictions of those models in Figures 8-10 in Appendix A.6.2 show that fine-tuned RoBERTa demonstrates some overconfidence with a high probability of incorrect labels, particularly when it comes to Fear, possibly related to a data imbalance.

6 Conclusion

This paper presents our contribution to the SemEval-2025's multi-label text-based emotion recognition task. In our work, we compare the emotion-detecting abilities of causal and non-causal models along with investigating multiple techniques such as curriculum learning, instruction and few-shot fine-tuning, as well as feature fusion with emotion scores (FFSA). Fine-tuned language-specific BERT-based models and XLM-RoBERTa-large gave the best results across multiple languages, outperforming baseline and generative models. For future work, we believe an interesting direction would be using data augmentation to address the lack of perceived emotion detection datasets. Additionally, we can explore improving the FFSA method using emotion lexicons.

Limitations

We acknowledge that our study on generative models for non-English languages is limited by basing some decisions solely on the English setup and applying them to other languages. Ideally, further studies should be conducted to identify the best experimental setup for each language. As for the test phase, we compared the two best approaches from the development phase on the test set for each selected language to report the final results. Another limitation is that emotion scores were computed only on training data samples, which may not fully capture real-world emotion dependencies.

Ethics Statement

As emotion recognition models heavily depend on the training data, biases presented in the datasets can be represented in the fine-tuned model version. Moreover, the possibility of misuse remains a significant concern, as the models could be used for manipulative purposes, such as generating targeted emotional responses or influencing public sentiment. We also acknowledge the computational resources required to work with LLMs, such as Llama-3, making it less accessible for lower-resource environments.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Anastasiia Demidova, Hanin Atwany, Nour Rabih, and Sanad Sha'ban. 2024. Arabic Train at NADI 2024 shared task: LLMs' ability to translate Arabic dialects into Modern Standard Arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 729–734, Bangkok, Thailand. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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*.

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. *Preprint*, arXiv:1907.11692.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. *arXiv e-prints*, arXiv:2308.08747.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. 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. *Preprint*, arXiv:2502.11926.

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.

Prashant Kumar Nag, Amit Bhagat, R. Vishnu Priya, and Deepak Kumar Khare. 2023. Emotional intelligence through artificial intelligence: Nlp and deep learning in the analysis of healthcare texts. In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, page 1–7. IEEE.

Irina Nikishina, Polina Chernomorchenko, Anastasiia Demidova, Alexander Panchenko, and Chris Biemann. 2023. Predicting terms in IS-a relations with pre-trained transformers. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 134–148, Nusa Dua, Bali. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Qwen. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jennifer Santoso, Takeshi Yamada, Shoji Makino, Kenkichi Ishizuka, and Takekatsu Hiramura. 2021. Speech emotion recognition based on attention weight correction using word-level confidence measure. In *Interspeech 2021*, pages 1947–1951.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17.

Zhiwei Yang, Hechang Chen, Jiawei Zhang, Jing Ma, and Yi Chang. 2020. Attention-based multi-level feature fusion for named entity recognition. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3594–3600. International Joint Conferences on Artificial Intelligence Organization. Main track.

Zhiwei Yang, Jing Ma, Hechang Chen, Jiawei Zhang, and Yi Chang. 2024. Context-aware attentive multilevel feature fusion for named entity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):973–984.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. DualGATs: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408, Toronto, Canada. Association for Computational Linguistics.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.

A Appendix

A.1 Data

A.1.1 Cross-lingual analysis

Table 4 provides analytical information across selected languages, indicating small sizes of datasets and extraordinal cases with long sentences (especially in Brazilian Portuguese and Swedish data). This may influence a model's performance due to the limited input size of a model.

Language	Size	L_{max}	L_{mean}
English	2768	450	78
Algerian Arabic	901	274	76
Moroccan Arabic	1608	444	77
German	2603	856	219
Nigerian-Pidgin	3728	279	111
Brazilian Portuguese	2226	2665	114
Mozambican Portuguese	1546	147	65
Russian	2679	609	62
Spanish	1996	191	53
Swedish	1187	3476	196

Table 4: Analytical information of training data among all selected languages. We present the number of samples per language (*Size*), as well as the sentence lengths in terms of the number of characters, showing the mean and max values across languages.

A.1.2 Emotional Combinations

The heat map in Figure 3 of such combinations distinctly illustrates co-occurrence rates for English training data. Some pairs of emotions occur more frequently than others, as indicated by their higher probability. This may reflect real-life tendencies, where the most commonly expressed emotions appear more often.

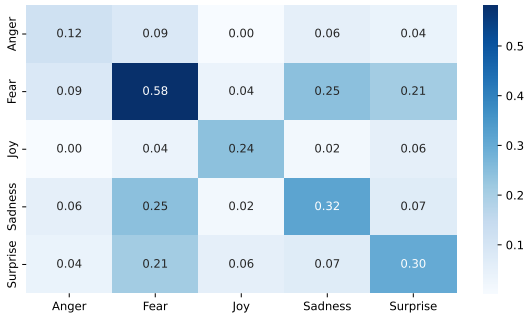


Figure 3: Heat map of pairwise probabilities for English train set.

English utterances with Neutral and Joy-only emotions, combinations of (Fear with Sadness), (Fear with Surprise), and (Fear with both Sadness and Surprise) demonstrate strong correlations.

A.1.3 Emotion Distribution

In addition, Figure 4, which shows the emotion distribution, confirms that the overall low probabilities of combinations involving Anger and Joy illustrate data imbalance. This may reflect emotional states that commonly co-occur in social media texts.

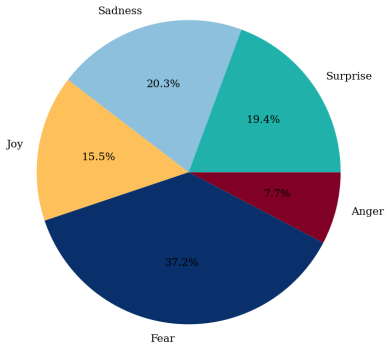


Figure 4: Emotion distribution in English Train dataset.

A.2 Hyper-parameters

A.2.1 Non-causal Models

Table 5 presents the tested hyper-parameter ranges. During experiments, we found the most optimal set of these hyper-parameters based on model performance. Figure 5 demonstrates the process of finding the optimal epoch number, as well as experiments with the HuggingFace Trainer[7], which outperformed our custom training function.

Hyper-parameter	From	To	Optimal
Epochs	1	20	10
Batch size	8	32	32
Learning rate	1e-6	3e-5	2e-5
Weight decay	1e-6	5e-6	1e-6

Table 5: Range of tuned hyper-parameters for RoBERTa-large in English setup.

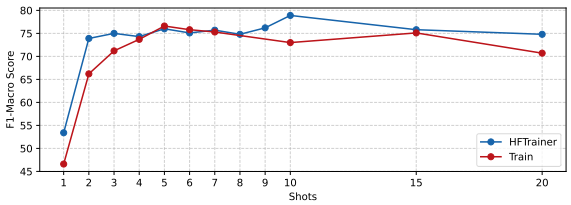


Figure 5: The plot of F1-Macro score results over the number of epochs of RoBERTa-large on the English development set using the custom training function (*Train*) and HuggingFace Trainer (*HFTrainer*).

[7] https://huggingface.co/docs/transformers/trainer

A.2.2 Causal Models

In Table 6, the demonstrated LoRA hyper-parameters allow us to regularize reducing memory usage by freezing most model weights and adapting all linear layers in a model.

Hyper-parameter	Value
LoRA α	16
dropout	0.1
r	64
bias	'none'
target modules	"all-linear"

Table 6: LoRA hyper-parameters for fine-tuning Llama-3-8B-Instruct in English setup.

During the inference process, we use certain generation parameters, shown in Table 7, for both fine-tuning and in-context learning approaches to control the model output. These parameters reduce creativity to ensure the model generates consistent responses while maintaining emotional context.

Parameter	Value
max_new_tokens	50
do_sample	True
temperature	0.1
top_p	0.6

Table 7: Parameters for the inference of Llama-3-8B-Instruct in English setup.

A.3 Postprocessing

To align emotions from the model's responses according to the basic six emotions, we use emotion mapping from Table 8. For the English setting, we map Disgust to Fear emotion.

Emotion	Variations
Anger	Anger, Annoyance, Disapproval
Disgust	Disgust
Fear	Fear, Nervousness
Joy	Joy, Amusement, Approval, Excitement, Gratitude, Love, Optimism, Relief, Pride, Admiration, Desire, Caring
Sadness	Sadness, Disappointment, Embarrassment, Grief, Remorse
Surprise	Surprise, Realization, Confusion, Curiosity

Table 8: Emotion mapping from Demszky et al. (2020).

A.4 In-Context Learning.

As shown in Figure 6, our prompt templates have an instruction format where we utilize special tokens for structuring. Also, we used complex and diverse examples from the training dataset presented in Table 9 in a few-shot setup.

#	Utterance	Labels
1	*"The cop tells him to have a nice day and walks away."*	Anger, Joy, Surprise
2	*"About 2 weeks ago I thought I pulled a muscle in my calf."*	Fear, Sadness
3	*"I got to babysit my grandson but my back hurt the next day."*	Joy, Sadness

Table 9: Selected representative samples for few-shot learning from the English Train dataset.

A.5 English Deep-dive Experiments

In Table 10, we provide results on all approaches in the English setup conducted using RoBERTa-large and Llama-3, where RoBERTa with a combination of fine-tuning and curriculum learning on preprocessed data shows the highest 81 F1-score.

Model	Approach	F1-Macro
RoBERTa	Preprocessing + CL	**81.0**
RoBERTa	FFSA	80.4
RoBERTa	CL	79.9
RoBERTa	Preprocessing	79.0
RoBERTa	Fine-tune	78.9
Llama-3	Prompt	73.0
Llama-3	Few-shot fine-tune	68.5
Llama-3	Instruction fine-tune	64.3

Table 10: Best results on the English development set, where *RoBERTa* - RoBERTa-large, *Llama-3* is Meta-Llama-3-8B-Instruct, *FFSA* - feature fusion with self-attention, *CL* - curriculum learning.

We found that preprocessing steps benefit the RoBERTa, which is optimized for clean and structured input such as Wikipedia[8] and BookCorpus (Zhu et al., 2015). In contrast, Llama-3 did not show better results on the preprocessed data compared to the original dataset. As a decoder, Llama-3 appears to be more robust to raw text variations because they are trained to handle natural instances.

A.6 Result Analysis

A.6.1 Confusion Matrices

For the English dataset, Figures 7a and 7b demonstrate confusion matrices of the two effective approaches, such as fine-tuning and feature fusion with self-attention. They represent a comparison between predicted and true labels, indicating a low

[8] https://dumps.wikimedia.org/

Input:
My friend told me: "*<Utterance>*".
What emotions from the list *[Anger, Fear, Joy, Sadness, Surprise]* did my friend convey to me in their own words?

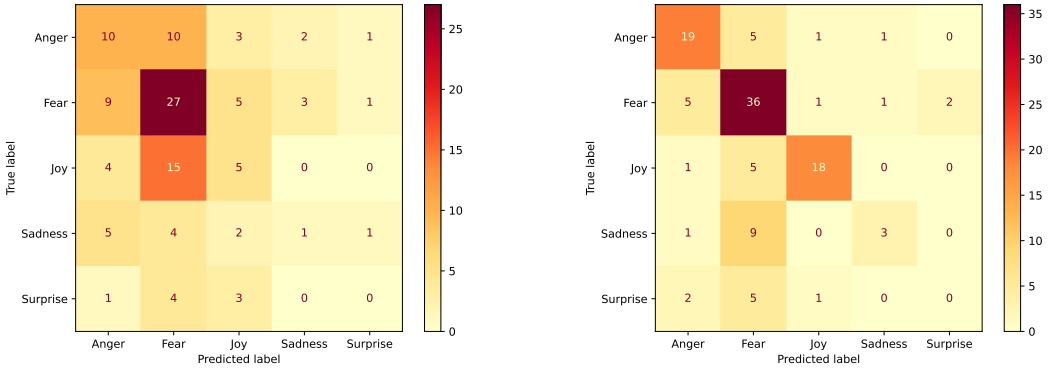
Output:
Emotions:

1

Input:
Your task is to predict the likely emotion(s) most people will think the speaker may be feeling in this speech: *<Utterance>*
[End of the speech]
Determine the most possible emotions from this list *[Anger, Fear, Joy, Sadness, Surprise, Neutral]*. If there are several emotions, write them separated by commas.
Output:
Emotion(s):

2

Figure 6: Prompt templates for the English setup.



(a) Fine-tuning approach

(b) Feature Fusion with Self-Attention

Figure 7: Confusion Matrices of RoBERTa for the English development set.

number of samples with Sadness and Surprise labels in the development set as well.

A.6.2 Error Analysis

Figures 8, 9, and 10 represent predicted probabilities for each label on the English development set. Here, high values indicated with blue colour reflect the overconfidence of a model, whereas low probabilities with red colour represent the underconfident model. Both of these cases indicate the need for calibration.

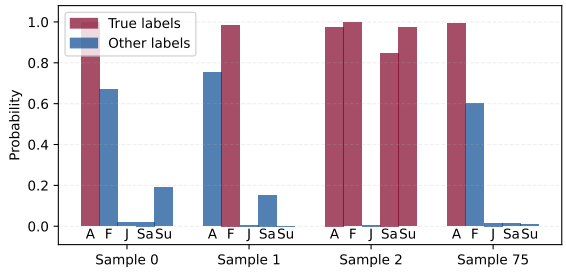


Figure 8: Examples of predicted probabilities of the fine-tuned RoBERTa on the development set (Anger (*A*), Fear (*F*), Joy (*J*), Sadness (*Sa*), Surprise (*Su*)).

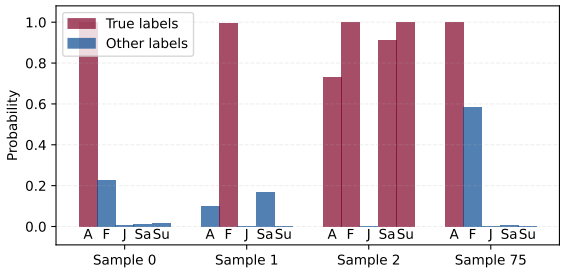


Figure 9: Examples of predicted probabilities of the fine-tuned RoBERTa with Feature Fusion and Self-Attention on the development set (Anger (*A*), Fear (*F*), Joy (*J*), Sadness (*Sa*), Surprise (*Su*)).

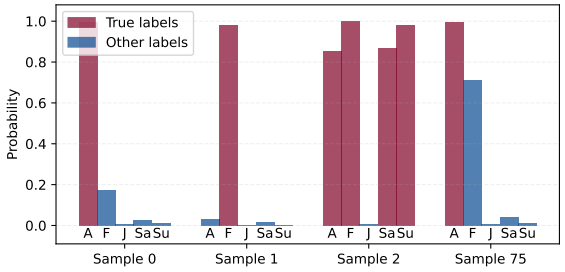


Figure 10: Examples of predicted probabilities of the fine-tuned RoBERTa with Curriculum Learning and Data Preprocessing (Anger (*A*), Fear (*F*), Joy (*J*), Sadness (*Sa*), Surprise (*Su*)).