Heimerdinger at SemEval-2025 Task 11: A Multi-Agent Framework for Perceived Emotion Detection in Multilingual Text

Zeliang Tong[1,♡], **Zhuojun Ding**[1,♡], **Yingjia Li**[2,♡]
[1] Huazhong University of Science and Technology, Wuhan, China
[2] Tsinghua University, Beijing, China
tongzeliang744@gmail.com, dingzj@hust.edu.cn, liyj23@mails.tsinghua.edu.cn

Abstract

This paper presents our system developed for the SemEval-2025 Task 11: Text-Based Emotion Detection (TBED) task, which aims to identify the emotions perceived by the majority of people from a speaker's short text. We introduce a multi-agent framework for emotion recognition, comprising two key agents: the Emotion Perception Profiler, which identifies emotions in text, and the Intensity Perception Profiler, which assesses the intensity of those emotions. We model the task using both generative and discriminative approaches, leveraging BERT series and large-scale generative language models (LLMs). A multi-system collaboration mechanism is employed to further enhance the accuracy, stability, and robustness. Additionally, we incorporate cross-lingual knowledge transfer to improve performance in diverse linguistic scenarios. Our method demonstrates superior results in emotion detection and intensity prediction across multiple subtasks, highlighting its effectiveness, especially in language adaptability. Our code is available at https://github.com/tongzeliang/Semeval2025

1 Introduction

Emotion perception is a complex and subtle process involving how individuals perceive, express, and interpret emotions (Canales and Martínez-Barco, 2014; Ghosal et al., 2021; Zhang et al., 2023). This paper focuses on the Text-Based Emotion Detection (TBED) task proposed in the SemEval-2025 Task 11 (Belay et al., 2025; Muhammad et al., 2025b), which aims to determine the emotion that the majority of people perceive the speaker to be experiencing based on a sentence or a short text fragment uttered by the speaker (along with further assessment of the intensity of emotions).

The TBED task concerns the emotion perceived by the majority of people rather than the speaker's actual emotion. For example, the sentence "I finally finished all my work today; this is fantastic" may evoke feelings of joy or satisfaction in most people. However, it does not rule out the possibility that the speaker's actual emotion might differ, perhaps due to some form of reverse expression. Furthermore, this task differs from the following emotion-related tasks: (1) reader emotions induced by the text (Rao et al., 2014) and (2) emotions of individuals mentioned in the text (Wegge and Klinger, 2023).

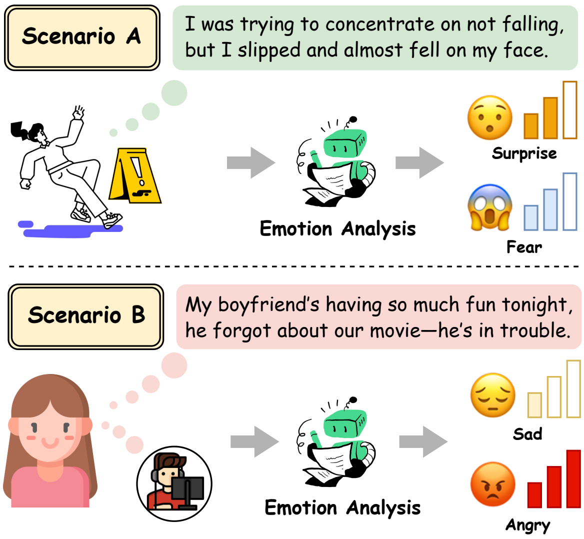

Scenario A
I was trying to concentrate on not falling, but I slipped and almost fell on my face.
Emotion Analysis
Surprise
Fear
Scenario B
My boyfriend's having so much fun tonight, he forgot about our movie—he's in trouble.
Emotion Analysis
Sad
Angry

Figure 1: Examples of TBED. The goal of this task is to recognize what emotion most people think the speaker might have felt given a sentence or a short text snippet uttered by the speaker.

In this paper, we propose an innovative emotion recognition framework based on multi-agent collaboration (Guo et al., 2024). Specifically, we design two key agents: (1) the Emotion Perception Profiler, used for identifying emotions in text, primarily applied to subtasks A and C, and (2) the Intensity Perception Profiler, which further assesses emotion intensity based on the output of the first agent, applied to subtask B. We model the task using both generative and discriminative paradigms, training the intelligent agents using BERT series (Devlin,

[♡] Contribute equal to this work.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 397–406
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2018; Liu, 2019) models and large-scale generative language models (LLMs) (Yang et al., 2024; Grattafiori et al., 2024; Guo et al., 2025). Given that different base models acquire different capabilities during the pre-training phase, including learned knowledge, linguistic proficiency, and task adaptation, we further introduce a multi-system collaboration mechanism. The intelligent agents constructed on unified base models are referred to as an agent system. By fusing prediction results from multiple systems, we enhance the accuracy, stability, and robustness of the results. Additionally, for subtask C, we introduce a cross-linguistic knowledge transfer mechanism, significantly improving the performance in cross-linguistic scenarios.

For the final evaluation, we select the models that performed best on the validation set. We test all languages involved in subtasks A and B. For subtask C, we only test languages that do not have any training or validation data during the validation phase to ensure the reliability of the results. Our approach achieves outstanding performance across all three subtasks, validating the effectiveness of our method, particularly in terms of language adaptability and emotion intensity prediction capabilities.

2 System Overview

Preliminary. Given a sentence or text snippet $W = \{w_1, w_2, ..., w_n\}$ uttered by the speaker, a predefined emotion set $E = \{e_1, e_2, ..., e_m\}$, the objective of the three subtasks is as follows:

- **Subtask A**: Predict the perceived emotion(s) $\mathcal{S} = \{e_k | e_k \in E\}$ of the speaker.
- **Subtask B**: Predict the perceived emotion(s) and their corresponding emotional intensity $\mathcal{P} = \{(e_k, i_k) | e_k \in E, i_k \in I\}$, where $I = \{\text{low}, \text{moderate}, \text{high}\}$, denotes three degrees of the emotion intensity.
- **Subtask C**: Predict the perceived emotion(s) $\mathcal{S} = \{e_k | e_k \in E\}$ of the speaker in an unseen target language, without relying on labeled training data in that language.

Framework. Our framework consists of two agents, ***Emotion Perception Profiler*** (EPP) and ***Intensity Perception Profiler*** (IPP). Specifically, as shown in Figure 2, EPP receives the text input and focuses on detecting the speaker's emotion, while IPP receives both the text input and the previously detected emotion information, taking the responsibility of determining the intensity of that emotion. In the following sections, we will present two distinct implementations of these agents: one based on BERT and the other based on Large Language Models (LLMs).

2.1 BERT-based Method

This approach leverages a Pretrained Language Model (PLM) model to obtain embedded representations for the corresponding features, framing the problem as either a multi-label or single-label classification task.

Emotion Perception Profiler. Firstly, we adopt XLM-Roberta as our PLM to generate contextual word representations by:

$$\begin{aligned}\mathbf{H} =& \text{XLM-Roberta}\left(\{w_1, w_2, \dots, w_n\}\right), \\ =& [\mathbf{h}_{\text{[CLS]}}, \mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n, \mathbf{h}_{\text{[SEP]}}].\end{aligned} \tag{1}$$

For each emotion $e_i \in E$, a dedicated binary classification head is defined as follows:

$$p_i = \text{Sigmoid}\left(\mathbf{w}_i^\top \mathbf{h}_{\text{[CLS]}} + \mathbf{b}_i\right). \tag{2}$$

Finally, the speaker is considered to exhibit the corresponding emotion if the probability p_i exceeds 0.5. The training loss is defined as:

$$\mathcal{L}_i = y_i \log p_i + (1 - y_i) \log(1 - p_i), \tag{3}$$

where $y_i \in \{0, 1\}$ is the ground truth label for the i-th emotion in the predefined set.

Intensity Perception Profiler. Similar to EPP, we utilize a PLM to encode the textual information along with the corresponding emotions that are to be assessed for intensity:

$$\begin{aligned}\mathbf{H} =& \text{XLM-Roberta}\left(\{w_1, w_2, \dots, w_n\}, e_k\right), \\ =& [\mathbf{h}_{\text{[CLS]}}, \mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n, \mathbf{h}_{\text{[SEP]}}, \mathbf{h}_{\text{emo}}].\end{aligned} \tag{4}$$

The representation of e_k is subsequently passed through a fully connected layer with a softmax activation function, producing probability distributions across all intensities:

$$\mathbf{p}_k = \text{Softmax}\left(\mathbf{W}^\top \mathbf{h}_{\text{emo}} + \mathbf{b}_k\right). \tag{5}$$

The training loss is formulated as the cross-entropy loss between the ground truth and the predicted label distributions, similar to formula 3.

2.2 LLM-based Method

Given the outstanding performance of LLMs across various domains, such as text classification and sentiment analysis, we also employ an LLM-based approach to implement the EPP and IPP, transforming the three subtasks into text generation tasks. We **combine fine-tuning and ICL** by utilizing LoRA as the parameter-efficient fine-tuning (PEFT) method to optimize the model and incorporating

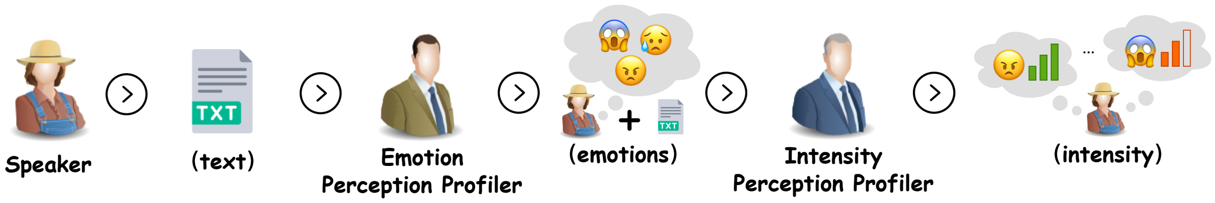
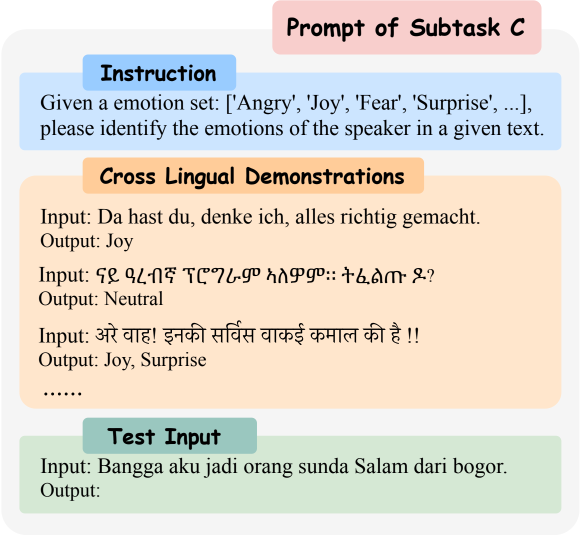


Figure 2: Our framework. The Emotion Perception Profiler analyzes the text output by the speaker to perceive the corresponding emotion, thereby addressing subtasks A and C. The Intensity Perception Profiler combines both emotional and textual information to assess the intensity of each emotion, completing subtask B.


Prompt of Subtask C

Instruction

Given a emotion set: ['Angry', 'Joy', 'Fear', 'Surprise', ...], please identify the emotions of the speaker in a given text.

Cross Lingual Demonstrations

Input: Da hast du, denke ich, alles richtig gemacht.
Output: Joy

Input: ናይ ዓረብኛ ፕሮግራም ኣለዎም። ትፈልጡ ዶ?
Output: Neutral

Input: अरे वाह! इनकी सर्विस वाकई कमाल की है !!
Output: Joy, Surprise

......

Test Input

Input: Bangga aku jadi orang sunda Salam dari bogor.
Output:

Figure 3: Examples of the prompt designed for subtask C. Specifically, the demonstrations used in ICL and the final test examples belong to different languages.

ICL demonstrations into the prompt, formulated as:

$$\hat{y} \sim P_{\mathcal{LLM}}(y \mid \mathcal{T}, x). \tag{6}$$

Here, $\mathcal{T} = \{\mathcal{I}, t(x_1, y_1), ..., t(x_k, y_k)\}$, where $\mathcal{I}$ represents the task instruction and t denotes the template of few-shot demonstrations for ICL. Notably, for subtask C, we introduce a novel **cross-lingual ICL** paradigm that leverages knowledge transfer from high-resource languages to enhance performance in low-resource settings where no target language training data is available.

Combine Fine-tuning and ICL. Conventionally, fine-tuning adapts a model to a specific task by explicitly adjusting its parameters, while in contrast, ICL performs the task through the prompting of examples. By incorporating ICL demonstrations into the prompt during the fine-tuning phase, the model can learn from both labeled data and contextual examples, enhancing its task understanding and generalization ability. Specifically, for subtasks A and B, we select the semantically closest top-k instances as ICL demonstrations by calculating the cosine distance between sentence embeddings.

Cross-lingual ICL. To address the zero-shot cross-lingual requirement of Subtask C, where no target-language training data is permitted, we propose a novel multilingual knowledge transfer framework through cross-lingual in-context learning, as shown in Figure 3. Our approach operates in two phases:

- **Fine-tuning:** During the PEFT phase via LoRA, we construct ICL demonstrations in $\mathcal{T}_m$ using multilingual examples $\{(x_s, y_s)\}$ from high-resource languages $s \in \mathcal{S}$ (e.g., English, Spanish), while maintaining the target language t exclusive from $\mathcal{T}_m$ exclusively for inference. The model is optimized to learn language-agnostic patterns through:

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[-\log P_{\mathcal{LLM}}(y \mid \mathcal{T}_{\text{m}}, x) \right] \tag{7}$$

  where $\mathcal{T}_{\text{m}}$ contains demonstrations from multiple high-resource source languages.

- **Inference:** At inference time for target language t, we retrieve semantically similar examples from other source languages using:

$$\text{sim}(x_t, x_s) = \cos(\mathbf{E}(x_t), \mathbf{E}(x_s)) \tag{8}$$

  where $\mathbf{E}$ denotes a multilingual sentence encoder (e.g., LaBSE). The top-k most relevant source-language examples $\{(x_{s_i}, y_{s_i})\}_{i=1}^{k}$ are injected into the prompt $\mathcal{T}_m$.

This dual-phase approach enables three key advantages: (1) *Cross-lingual capability activation* by leveraging ICL demonstrations from high-resource languages to stimulate the model's latent understanding of the target low-resource language, (2) *Cross-lingual semantic bridging* via contrastive-aligned multilingual embeddings, and (3) *Zero-shot generalization* without violating the target-language data constraint. In the Experiment section, we further elaborate on and summarize the selection strategy for the source high-resource language derived from empirical evidence.

Method		Chinese(chn)		German(deu)		English(eng)		Spanish(esp)		Portuguese(ptbr)		Russian(rus)	
		A	B	A	B	A	B	A	B	A	B	A	B
XLM-Roberta		59.10	**70.26**	65.00	67.08	72.80	**80.06**	79.40	75.16	57.20	56.88	85.90	88.66
LLMs	LLaMa3.1-8B	**70.62**	69.98	67.80	**67.91**	80.09	79.82	81.67	75.28	53.70	56.00	**89.77**	87.95
	Deepseek-7B	66.30	69.11	62.30	64.78	78.70	78.33	77.30	74.94	51.40	55.67	85.20	88.24
	Qwen2.5 -7B	67.30	69.72	**70.90**	67.19	79.10	78.30	81.90	**75.62**	**61.60**	**56.94**	87.10	**88.70**
	Mistralv0.3-7B	70.10	68.79	65.20	66.72	**82.60**	77.92	**82.08**	74.19	52.20	56.99	88.10	87.48
	Gemma2-9B	62.10	–	60.80	–	81.20	–	75.10	–	53.10	–	79.20	–

Table 1: Experimental results on the validation set of subtasks A and B. Average F1-Macro scores are reported for subtask A. Pearson correlation scores are reported for subtask B. The best performance scores are in **bold**.

	Source languages	Indonesian(ind)	Javanese(jav)	isiXhosa(xho)	isiZulu(zul)
Qwen2.5-7B	Portuguese(Brazilian; ptbr)	37.00	30.32	8.48	4.09
	Russian(rus)	39.49	31.79	10.61	11.89
	German(deu)	36.06	26.22	5.43	4.28
	Spanish(esp)	47.86	**38.92**	**16.87**	**16.15**
	Swahili(swa)	34.95	28.89	–	–
	Sundanese(sun)	**51.28**	36.68	–	–
	Chinese(chn)	37.70	28.41	6.41	4.61

Table 2: Experimental results on the validation set of subtask C. Average F1-Macro scores are reported. The best performance scores are in **bold**.

	Source languages	ind	jav
Qwen2.5-7B	all	51.39	37.13
	sun+esp	54.39	**41.46**
	sun+esp+rus+swa	49.43	35.50
	sun+esp+rus	**55.06**	38.40
	all w/o swa	45.99	39.15
	all w/o deu	47.20	40.34
	all w/o sun	46.67	37.54
	all w/o ptbr	46.45	39.98

Table 3: Experimental results on the validation set of subtask C. In contrast to results in Table 2, we employ a diverse set of source languages for model training.

3 Experiments

3.1 Setup

Models and Metrics. For discriminative modeling, we employ the XLM-Roberta-Large (Liu, 2019) as the encoder and employ linear lays as classifiers. For generative modeling, we utilize several prominent large language models (LLMs), including LLaMa3.1-8B (Dubey et al., 2024), Deepseek-7B (Bi et al., 2024), Qwen2.5-7B (Yang et al., 2024), Mistralv0.3-7B (Jiang et al., 2023), and Gemma2-9B (Team et al., 2024). We adopt LoRA (Hu et al., 2021) as our parameter-efficient fine-tuning method.

Metrics. The evaluation metric for subtasks A and C is the F1-macro based on the predicted and gold labels. Subtask B employs the Pearson correlation between the predicted labels and the gold ones for evaluation.

3.2 Main Results

During the validation phase, we evaluate six languages for subtasks A and B, with the results presented in Tables 1. Despite these models being pre-trained on multiple languages, disparities still exist. Overall, the performance of LLaMA and Qwen is relatively superior. Additionally, the effectiveness of LLMs generally surpasses that of BERT-based (smaller) models (SLMs).

Notably, for subtask A, the superiority of LLMs is more pronounced. However, for subtask B, there is no significant difference between the two. This discrepancy may be attributed to the different evaluation metrics employed for subtasks A and B. This observation provides valuable insights, suggesting that it may be beneficial to integrate both large and small models to leverage their respective strengths for multi-agent collaboration. Furthermore, due to the adoption of a multi-step mode, a potential issue of exposure bias may arise. However, our experiments revealed that end-to-end modeling approaches yielded inferior results, which may be related to the complexity of the tasks. We also

Languages	Baselines					Ours
	Qwen2.5-72B	Dolly-v2-12B	Llama-3.3-70B	Mixtral-8x7B	Deepseek-R1-70B	
Afrikaans(afr)	60.18	23.58	**61.28**	53.69	43.66	57.37$_{(\downarrow 3.91)}$
Algerian Arabic(arq)	37.78	38.59	55.75	45.29	50.87	**59.48**$_{(\uparrow 3.73)}$
Moroccan Arabic(ary)	52.76	24.27	44.96	35.07	47.21	**58.19**$_{(\uparrow 5.43)}$
Chinese(chn)	55.23	27.52	53.36	44.91	53.45	**68.00**$_{(\uparrow 12.77)}$
German(deu)	59.17	26.86	56.99	51.20	54.26	**70.64**$_{(\uparrow 11.47)}$
English(eng)	55.72	42.60	65.58	58.12	56.99	**80.40**$_{(\uparrow 14.82)}$
Spanish(esp)	72.33	36.41	61.27	65.72	73.29	**83.83**$_{(\uparrow 10.54)}$
Hausa(hau)	43.79	29.43	50.91	40.40	51.91	**61.54**$_{(\uparrow 9.63)}$
Hindi(hin)	79.73	27.59	60.59	62.19	76.91	**89.56**$_{(\uparrow 9.83)}$
Igbo(ibo)	37.40	24.31	33.18	31.90	32.85	**54.12**$_{(\uparrow 16.72)}$
Kinyarwanda(kin)	31.96	19.73	34.36	26.35	32.52	**44.51**$_{(\uparrow 10.15)}$
Marathi(mar)	74.58	25.69	67.40	50.36	76.68	**87.74**$_{(\uparrow 11.06)}$
Nigerian-Pidgin(pcm)	38.66	34.41	48.67	45.61	45.00	**60.45**$_{(\uparrow 11.78)}$
Portuguese(Brazilian; ptbr)	51.60	25.90	45.03	41.64	51.49	**62.46**$_{(\uparrow 10.86)}$
Portuguese(Mozambican; ptmz)	40.44	16.70	34.06	36.52	39.58	**50.72**$_{(\uparrow 10.28)}$
Romanian(ron)	68.18	43.58	71.28	68.51	65.02	**74.60**$_{(\uparrow 3.32)}$
Russian(rus)	73.08	29.72	62.61	61.72	76.97	**90.08**$_{(\uparrow 13.11)}$
Sundanese(sun)	42.67	32.20	46.33	42.10	44.61	**48.16**$_{(\uparrow 1.83)}$
Swahili(swa)	27.36	17.63	29.47	26.51	**33.27**	30.23$_{(\downarrow 3.04)}$
Swedish(swe)	48.89	21.79	50.26	48.61	44.60	**58.15**$_{(\uparrow 7.89)}$
Tatar(tat)	51.58	25.12	49.84	39.44	53.86	**75.43**$_{(\uparrow 21.57)}$
Ukrainian(ukr)	54.76	17.16	42.34	40.15	51.19	**63.70**$_{(\uparrow 8.94)}$
Emakhuwa(vmw)	20.41	16.03	18.96	19.00	19.09	**22.17**$_{(\uparrow 1.76)}$
Yoruba(yor)	24.99	16.00	23.70	19.67	27.44	**39.19**$_{(\uparrow 11.75)}$
Avg	50.14	26.78	48.67	43.95	50.11	**62.11**$_{(\uparrow 11.97)}$

Table 4: Experimental results on the test set of subtask A. Average F1-Macro scores are reported. The best performance scores are in **bold**.

Languages	Baselines					Ours
	Qwen2.5-72B	Dolly-v2-12B	Llama-3.3-70B	Mixtral-8x7B	Deepseek-R1-70B	
Algerian Arabic(arq)	29.54	3.80	36.29	31.05	36.37	**50.67**$_{(\uparrow 14.30)}$
Chinese(chn)	46.17	8.11	51.86	46.52	48.57	**67.09**$_{(\uparrow 15.23)}$
German(deu)	43.30	7.43	53.46	47.60	54.78	**72.29**$_{(\uparrow 17.51)}$
English(eng)	55.99	13.35	44.14	55.26	48.08	**79.34**$_{(\uparrow 23.35)}$
Spanish(esp)	51.11	10.49	51.64	55.54	60.74	**78.75**$_{(\uparrow 18.01)}$
Hausa(hau)	27.00	6.43	39.16	25.84	38.85	**61.76**$_{(\uparrow 22.60)}$
Portuguese(Brazilian; ptbr)	38.20	9.02	40.90	39.17	46.72	**65.06**$_{(\uparrow 18.34)}$
Romanian(ron)	55.48	12.62	45.87	57.07	57.69	**65.71**$_{(\uparrow 8.02)}$
Russian(rus)	58.25	13.96	57.56	56.01	62.28	**88.34**$_{(\uparrow 26.06)}$
Ukrainian(ukr)	37.74	6.04	36.99	38.74	43.54	**60.34**$_{(\uparrow 16.80)}$
Avg	44.28	9.13	45.79	45.28	49.76	**68.94**$_{(\uparrow 19.18)}$

Table 5: Experimental results on the test set of subtask B. Pearson correlation scores are reported. The best performance scores are in **bold**.

Languages	Baselines					Ours	
	Qwen2.5-72B	Dolly-v2-12B	Llama-3.3-70B	Mixtral-8x7B	Deepseek-R1-70B	F1	RANK
Indonesian(ind)	57.29	36.61	39.20	54.37	49.51	**60.90**	3rd
Javanese(jav)	**50.47**	36.18	41.88	48.37	43.05	43.86	1st
isiXhosa(xho)	29.56	24.12	**30.79**	22.92	29.08	26.03	3rd
isiZulu(zul)	22.03	14.72	21.48	20.38	20.38	**22.56**	3rd
Avg	**39.84**	27.91	33.34	36.51	35.51	38.34	-

Table 6: Experimental results on the test set of subtask C. Average F1-Macro scores are reported. The best performance scores are in **bold**.

Method		English	German	Portuguese
LLaMa3.1-8B	ours	**80.09**	**67.80**	**53.70**
	-w/o ICL	78.35	67.01	49.16
Qwen2.5-7B	ours	**79.10**	**70.90**	**61.60**
	-w/o ICL	77.92	68.34	59.77
Mistral0.3-7B	ours	**82.60**	**65.20**	**52.20**
	-w/o ICL	79.88	63.50	50.36

Table 7: Ablation results of subtask A. "w/o" ICL denotes the removal of demonstrations from the prompt. The best results are highlighted in bold.

Method		English	German	Portuguese
LLaMa3.1-8B	ours	**79.82**	**67.91**	**56.00**
	-w/o ICL	76.34	65.48	53.72
	-w IPP$^+$	77.20	66.00	54.26
	-wCodeIPP$^+$	78.47	67.63	55.01
Qwen2.5-7B	ours	**78.30**	**67.19**	**56.94**
	-w/o ICL	77.92	68.34	59.77
	-wIPP$^+$	76.47	65.23	54.10
	-wCodeIPP$^+$	76.99	66.31	54.99
Mistral0.3-7B	ours	**77.92**	**66.72**	**56.99**
	-w/o ICL	76.23	63.34	54.87
	-wIPP$^+$	75.31	63.03	54.30
	-w CodeIPP$^+$	76.36	62.69	55.07

Table 8: Ablation results of subtask B. "IPP$^+$" refers to the approach where a single agent directly performs both emotion prediction and intensity detection tasks. "CodeIPP$^+$" extends IPP$^+$ by utilizing a code-style prompt, modeling emotion and intensity as structured pairs, and applying the corresponding LLM code version. The best results are highlighted in bold, and the second-best results are underlined.

preliminarily verified that code-style prompts can enhance the performance of end-to-end modeling approaches. Table 2 and 3 illustrate the validation set performance for subtask C. We trained models using different source languages and tested them on four target languages. We conducted experiments under two distinct scenarios: one involving training the model using a single source language exclusively, and the other incorporating multiple languages for model training. The results indicate that source languages linguistically closer to the target languages enhance the model performance.

During the testing phase, we compare our approach with official LLM baselines (Muhammad et al., 2025a; Belay et al., 2025), including Qwen2.5-72B, Dolly-v2-12B, Llama-3.3-70B, Mixtral-8x7B, and DeepSeek-R1-70B. The results for the three subtasks are presented in Tables 4, 5 and 6, respectively. Even though these baselines employed significantly larger models, our approach demonstrated superior performance.

3.3 Ablation Study

We conduct a series of ablation experiments to systematically evaluate the contribution of each proposed optimization, as shown in Tables 7 and 8. The results reveal several key insights: (1) Incorporating ICL demonstrations during fine-tuning improves model performance. This can be attributed to the dual role of ICL examples in providing explicit task-specific guidance and enhancing the model's ability to generalize from contextual patterns, thereby bridging the gap between pre-trained knowledge and downstream task requirements. (2) The single-agent IPP$^+$ setup underperforms compared to the multi-agent approach, highlighting the importance of task decomposition and collaborative reasoning. The multi-agent system likely benefits from specialized handling of subtasks, enabling more robust decision-making through inter-agent interactions. (3) Replacing natural language prompts with code-style formatting further enhances performance. This improvement may stem from the structured nature of code-style prompts, which enforce stricter syntactic and semantic constraints, reducing ambiguity and aligning more effectively with the model's pre-trained capabilities in code comprehension and structured reasoning. Collectively, these findings demonstrate that each optimization contributes meaningfully to the overall performance, validating the design choices of our framework.

4 Conclusion

In this paper, we propose a multi-agent framework for the Text-Based Emotion Detection (TBED) task in the SemEval-2025 Task 11. Our system, comprising the Emotion Perception Profiler and Intensity Perception Profiler, demonstrates superior accuracy, stability, and robustness across multiple subtasks. The cross-lingual knowledge transfer and in-context learning strategies effectively address challenges in low-resource languages, enhancing language adaptability. Extensive experiments demonstrate the effectiveness of our approach without the need for costly pre-training. Future work will explore further optimizations in diverse linguistic and emotional contexts.

Limitations

We acknowledge the following limitations of our work: (1) For subtask B, the sequential process of first determining sentiment and then predicting intensity inevitably introduces exposure bias. (2) The necessity of employing multiple collaborative agents incurs additional computational and storage overhead. (3) The absence of supplementary pre-training limits the model's adaptability to languages less encountered during its initial pre-training phase.

References

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Monali Bordoloi and Saroj Kumar Biswas. 2023. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial intelligence review*, 56(11):12505–12560.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

T Guo, X Chen, Y Wang, R Chang, S Pei, NV Chawla, O Wiest, and X Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*. IJCAI; Cornell arxiv.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chen Lin, Xinyi Liu, Guipeng Xv, and Hui Li. 2021. Mitigating sentiment bias for recommender systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 31–40.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. 2022. Dialogueein: Emotion interaction network for dialogue affective analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 684–693.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

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. *arXiv 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.

Ziqi Qiu, Jianxing Yu, Yufeng Zhang, Hanjiang Lai, Yanghui Rao, Qinliang Su, and Jian Yin. 2025. Detecting emotional incongruity of sarcasm by commonsense reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9062–9073.

Yanghui Rao, Qing Li, Liu Wenyin, Qingyuan Wu, and Xiaojun Quan. 2014. Affective topic model for social emotion detection. *Neural Networks*, 58:29–37.

Linfeng Song, Chunlei Xin, Shaopeng Lai, Ante Wang, Jinsong Su, and Kun Xu. 2022. Casa: Conversational aspect sentiment analysis for dialogue understanding. *Journal of Artificial Intelligence Research*, 73:511–533.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. 2023a. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. 2023b. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Maximilian Wegge and Roman Klinger. 2023. Automatic emotion experiencer recognition. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 1–7.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

Yice Zhang, Jie Zeng, Weiming Hu, Ziyi Wang, Shiwei Chen, and Ruifeng Xu. 2024. Self-training with pseudo-label scorer for aspect sentiment quad prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11862–11875.

You Zhang, Jin Wang, and Xuejie Zhang. 2021. Conciseness is better: Recurrent attention lstm model for document-level sentiment analysis. *Neurocomputing*, 462:101–112.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Wenting Zhao, Ye Liu, Yao Wan, Yibo Wang, Qingyang Wu, Zhongfen Deng, Jiangshu Du, Shuaiqi Liu, Yunlong Xu, and S Yu Philip. 2024. knn-icl: Compositional task-oriented parsing generalization with nearest neighbor in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 326–337.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

A Experimental Details

Datasets Table 9 presents the distribution of training data across various subtasks. Subtask A encompasses datasets in 27 languages, with a relatively balanced distribution. Each language has approximately 2,300 training samples on average. Subtask B includes datasets in 11 languages, with an average of 2,200 training samples per language. Subtask C evaluates the cross-lingual performance. The complete test dataset comprises 32 languages. However, most of these languages are included in subtask A. Therefore, we focus on languages that lack any training data, totaling four languages. Subtasks A and C require the identification of sentiment categories, including joy, sadness, fear, anger, surprise, and disgust (with some languages involving only five categories). Subtask B further assesses the intensity of the identified emotions, categorizing them into low, moderate, and high degrees. For more details, refer to Muhammad et al. (2025a); Belay et al. (2025).

Implementation Details For Bert models, we set the learning rates for the encoder and the classifier to 2e-5 and 1e-3, respectively. The batch size was set to 16, and the model was trained for 3 epochs using the AdamW optimizer. For LLMs, the training configuration consists of a learning rate of 1e-4, 3 epochs, a batch size of 2, bf16 mixed precision, and a maximum sequence length of 2048 tokens. Additionally, the rank of LoRA fine-tuning is set to 16, with a warmup ratio of 0.1. All implementations are conducted within the PyTorch framework, utilizing NVIDIA 4090 GPUs for computation. The number of demonstrations for ICL is set to 10.

B Supplementary Experiments

B.1 Model Comparision

To further investigate the nuanced performance variations of different models across languages and emotion categories, we conducted the systematic experiments outlined in Figure 4. Our analysis yields three principal findings:

(1) **Emotion-specific performance divergence**: The detection accuracy and intensity quantification efficacy exhibit significant disparities across emotion categories. We hypothesize that this phenomenon stems from differential boundary clarity in emotional intensity gradations. For instance, the superior detection of “Joy” across all languages contrasts with the suboptimal performance on “Disgust”, potentially attributable to the inherent ambiguity in disgust intensity demarcation and its infrequent contextual manifestations in training corpora. This aligns with Ekman’s (Ekman, 1992) basic emotion theory, where primary emotions exhibit more prototypical expressions, while complex emotions like disgust demonstrate higher cultural dependency. Joy-related lexicons show higher cross-lingual isomorphism in intensity scales (e.g., “delighted” vs. “content”) compared to disgust’s binary conceptualization (“disgusted” vs. “non-disgusted”) in most languages.

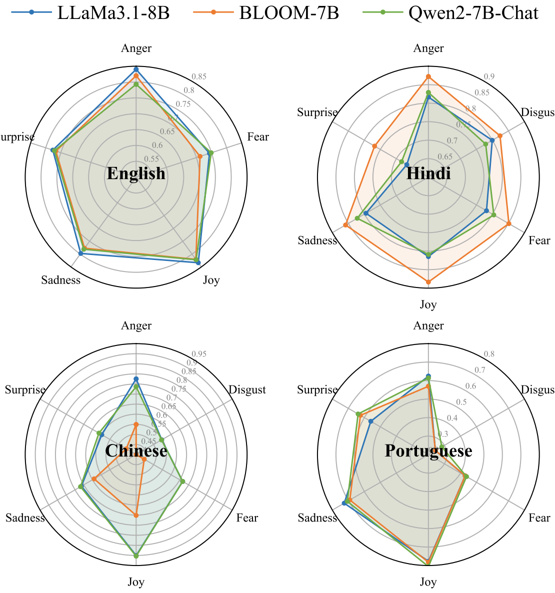


Figure 4: The Fine-grained performance of Subtask B. We compared the performance of three LLMs across four languages: English, Hindi, Chinese, and Portuguese. Each axis corresponds to a specific emotional category, with the performance of the models represented as follows: LLaMa3.1-8B (blue), BLOOM-7B (orange), and Qwen2-7B-Chat (green).

(2) **High-resource language performance plateau**: Model disparities remain statistically insignificant for linguistically well-resourced languages such as English and Portuguese. This convergence in performance across models suggests a saturation effect, where the optimization of model performance reaches a plateau in languages with abundant training data. Such saturation implies that, for these linguistically well-resourced languages, the need for careful selection between different LLMs may be less critical, as all models are likely to perform at a similar level of effectiveness.

(3) **Low-resource language sensitivity**: Conversely, marked performance discrepancies emerge in Hindi (BLOOM vs. LLaMa: F1=0.841 vs.

Subtask	Training Data Distribution	#Languages	#Avg
A	afr(1.9%), amh(5.5%), arq(1.4%), ary(2.5%), chn(4.1%), deu(4.0%), eng(4.3%), esp(3.1%), hau(3.3%), hin(3.9%), ibo(4.4%), kin(3.8%), mar(3.7%), orm(5.3%), pcm(5.7%), ptbr(3.4%), ptmz(2.4%), ron(1.9%), rus(4.1%), som(5.2%), sun(1.4%), swa(5.1%), swe(1.8%), tat(1.5%), tir(5.7%), ukr(3.8%), vmw(2.4%), yor(4.6%)	28	2.3k
B	amh(14.1%), arq(3.6%), chn(10.5%), deu(10.3%), eng(11.0%), esp(7.9%), hau(8.5%), ptbr(8.8%), ron(4.9%), rus(10.6%), ukr(9.8%)	11	2.2k
C	ind, jav, xho, zul	4	*Not available*

Table 9: Statistics of the training data. We show the proportion of training data in each language, the number of languages, and the average amount of data in each dataset. The training set for Subtask C is not available.

0.792) and Chinese (Qwen vs. BLOOM: F1=0.697 vs. 0.541). This underscores the importance of strategically selecting LLMs based on language-specific architectural optimization and the representativeness of training data, especially when dealing with under-resourced languages. The superior performance of BLOOM in Hindi potentially reflects its enhanced morphological processing capabilities, which are particularly well-suited for agglutinative languages like Hindi. This highlights the need for models that are not only trained on diverse multilingual datasets but also optimized to handle the unique syntactic and morphological characteristics of specific languages.

C Related Works

Sentiment Analysis. Sentiment analysis is a multifaceted area that seeks to understand how language conveys and perceives emotions. It can be divided into three levels: Document-level analysis (Zhang et al., 2021) captures the overall emotional tone of a text, useful for tasks like sentiment analysis in reviews. Sentence-level analysis (Bordoloi and Biswas, 2023) focuses on emotions within individual sentences, often applied in more detailed sentiment classification. Aspect-based Sentiment Analysis (ABSA) (Zhang et al., 2024) identifies emotions related to specific features or aspects, which is especially valuable in opinion mining, where emotions towards different components of a product or service need to be assessed separately. All three granularities have rich downstream applications, such as sarcasm detection (Qiu et al., 2025), dialogue system (Song et al., 2022; Liu et al., 2022), or recommendation system (Lin et al., 2021). In this paper, we focus primarily on sentence-level multi-label emotion classification tasks.

In-context Learning. In-context learning refers to the ability of models, especially LLMs, to adapt to a task by conditioning on a few examples or instructions provided within the input without requiring explicit retraining (Brown et al., 2020; Zhang et al., 2022; Wei et al., 2023; Zhao et al., 2024). Early work on this concept emphasized its potential in tasks such as few-shot learning, where models demonstrated impressive performance by simply leveraging context from examples embedded in the prompt (Zhao et al., 2021; Lu et al., 2022). Subsequent studies have explored how models can generalize across various tasks, including question answering and text generation, by relying on in-context examples (Wang et al., 2023b; Hendel et al., 2023; Wang et al., 2023a). The flexibility of in-context learning has made it a promising approach for tasks with limited labeled data or dynamic, context-sensitive applications. This paper proposes a novel cross-lingual ICL framework to enhance LLMs' adaptability for low-resource languages by strategically leveraging cross-lingual knowledge transfer through contextual demonstrations.