Cross-Lingual Emotion Recognition in Balinese Text using Multilingual-LLMs under Peer-Collaborations Settings

Putu Kussa Laksana Utama, Jilles Steeve Dibangoye, Tsegaye Misikir Tashu
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence,
University of Groningen
Correspondence: p.k.laksana.utama@rug.nl

Abstract

Cross-Lingual Emotion Recognition (CLER) remains a formidable challenge for ultra-low-resource languages like Balinese due to the scarcity of high-quality annotated data and the performance limitations of traditional multilingual models. This study addresses these gaps through two primary contributions. First, we present a newly created multi-label Balinese emotion dataset annotated by a panel of experts in Balinese linguistics and psychology. Second, we propose the Multi-Agent Peer Collaboration (MAPC) framework, which transforms the multi-label classification problem into a series of independent binary tasks to leverage the collaborative reasoning of Large Language Models (LLMs). We evaluated the framework against the LaBSE multilingual model and three LLMs of varying scales under zero-shot and few-shot settings using the Macro-F1 measure. The experimental results showed that LLMs significantly outperform traditional Pre-trained Language Models (PLMs). MAPC achieved an overall macro F_1-score of 63.95, which was higher than the individual baselines in both zero-shot and few-shot settings. Analysis shows that while some models exhibit sensitivity to few-shot prompting in low-resource contexts, the MAPC review and revision process consistently improves individual reasoning and provides a more accurate final classification.

1 Introduction

Cross-lingual emotion recognition (CLER) is a challenging task in natural language processing due to the complexity of emotion expression across languages and cultures. Monolingual emotion recognition benefits from rich, annotated corpora and language-specific features, while CLER faces data scarcity for many language pairs, vocabulary differences between languages, and language register shifts between training and target corpora. Bridging these gaps requires methods for transferring emotional knowledge and capturing language-agnostic representations. Recent works have tackled this by leveraging large pre-trained language models (PLMs) (Hassan et al., 2022; AliNadia et al., 2023), which use multilingual capabilities for zero-shot or few-shot emotion transfer.

Despite the success of PLMs in leveraging multilingual knowledge for cross-lingual tasks, a significant challenge persists, particularly in the low-resource language domain. Several studies (Nie et al., 2023; Hangya et al., 2022) have reported that the PLM approach remains suboptimal in low-resource language tasks, particularly for unseen languages, showing relatively low performance scores.

Furthermore, Hangya et al. (2022) mentioned that this degradation was attributed to several factors. First, the utilized BERT-based models (mBERT and eBERT) provide low-quality initial contextualized representations for these languages, which undermines every subsequent step. Because the models start with poor cross-lingual embeddings, the unsupervised mining of word translation pairs becomes ineffective. Second, unseen languages typically have limited monolingual data, causing subword tokenizers to oversplit words and further reducing the number of tokens with reliable contextualized representations. Finally, the initial model quality is so low that meaningful improvement requires a stronger training signal than unsupervised mining can provide; supervised approaches using parallel data yield substantially better results.

To address these limitations, we leveraged the current generation of LLMs, which offers a stronger foundation for cross-lingual tasks. LLMs are trained on vastly larger and more diverse multilingual datasets, resulting in a broader language coverage and more robust, high-quality representations, particularly for languages traditionally considered low-resource. As a result, they offer clear advantages over PLMs. Recent studies in cross-

lingual tasks (Wang et al., 2023; Huang et al., 2023; Zhang et al., 2024) and particularly in emotion-recognition (Xue et al., 2025a), demonstrate the promising result of LLM performance.

Despite the abundance of evidence demonstrating the significant performance gains of LLMs across various low-resource languages, there remains an need to investigate their capabilities in the cross-lingual emotion recognition task for other, currently under-explored languages, such as Balinese [1]. The importance of this focus stems not only from the linguistic and cultural significance of Balinese but also from the opportunity to challenge the generalization of state-of-the-art LLMs in highly constrained settings. To the best of our knowledge, there has been no prior work dedicated to emotion recognition in Balinese text using a cross-lingual approach. We hypothesize that success in this setting will create opportunities for adapting LLMs to other ultra-low-resource languages.

To address the aforementioned challenges, first, we construct the first-ever Balinese emotion dataset, which serves as our evaluation benchmark, since no such corpus is currently available for this low-resource language. In addition, we release our datasets to the public [2] to facilitate and encourage further investigations of emotion recognition in low-resource language settings. Second, in a multi-label, cross-lingual setting, we investigate the challenging scenario of identifying whether one or more emotions co-occur within a single text instance (i.e., multi-label classification). Third, to facilitate knowledge transfer and minimize the performance gap, we leverage the LLM's multilingual capacity by strategically employing related Austronesian languages, specifically Indonesian, Javanese, and Sundanese, as few-shot examples within the prompt context. These related language emotion datasets are sourced from the publicly available BRIGHTER dataset (Muhammad et al., 2025). Finally, we prompt the advanced LLM to classify the emotion(s) in the target Balinese text.

Building upon the robust foundation provided by LLMs, we argue that a single, monolithic LLM agent is insufficient to fully address the challenges of CLER in ultra-low-resource settings. This assumption arises from several recent works (Chen et al., 2025; Chang et al., 2024) that demonstrate different LLMs possess inherently distinct language capabilities. We recognize that different LLMs exhibit complementary strengths: some capture fine-grained linguistic structures, whereas others more effectively infer subtle semantic linguistic expression through broader contextual understanding.

To harness these complementary strengths, we propose a joint collaboration mechanism implemented in peer-collaboration settings involving more than two distinct LLM agents. This multi-agent framework allows agents to share knowledge, critique predictions, and reach a consensus, simulating a peer review process to enhance the emotion recognition in the target Balinese language. The contributions of this paper are as follows:

1. We address the foundational data scarcity problem by constructing and publicly releasing the first-ever emotion dataset for the Balinese language.

2. We introduce a novel, knowledge-intensive framework by adapting state-of-the-art LLMs to the CLER task using a structured peer-collaboration setting. This mechanism leverages the complementary strengths of multiple LLM agents to collectively enhance the performance of cross-lingual knowledge transfer from related languages (Indonesian, Javanese, Sundanese) to Balinese.

The remainder of this paper is organized as follows: Section 2 discusses theories of emotion, the fundamentals of the cross-lingual task in low-resource settings, multi-label classification, and related cross-lingual models. Section 3 details the dataset construction and the proposed Multi-Agent Peer Collaboration method. Section 4 presents the experimental results, discusses the findings, and section 5 outlines the limitations. Finally, Section 6 concludes the paper with a summary of our work.

2 Background and Related Work

Human emotional expression is primarily represented across two widely accepted theoretical frameworks: the categorical model and the dimensional model. The categorical model, proposed by Ekman (1992), posits that a limited set of universally distinct emotions exists, often referred to as basic emotions. Ekman formulated that

[1] Balinese is an Austronesian language spoken mainly in Bali island, Indonesia.

[2] The dataset is available at https://huggingface.co/datasets/kyo21/BaliEmoV1.

these fundamental emotions typically comprise six categories: joy, sadness, anger, fear, disgust, and surprise. However, Plutchik (1991) later expanded this set to eight core categories, incorporating concepts like anticipation and trust alongside Ekman's original six. To manage the scope and complexity of the current CLER task, which targets low-resource Austronesian languages, we adopt the categorical model based on Ekman's six basic emotions. This selection is supported by the fact that the majority of existing, publicly available cross-lingual emotion datasets pertaining to Austronesian languages utilize only these six basic emotional categories (Muljono et al., 2016; Yohanes et al., 2023; Putra et al., 2020), and also focusing on the six basic emotions mitigates the computational burden associated with multi-label classification across a larger, more complex set of labels (Bi and Kwok, 2013).

Emotion recognition from text is often modeled as a multi-label classification problem, where a single text instance can convey multiple co-occurring emotions (Ameer et al., 2023; Liu et al., 2023). This setting presents critical challenges such as label imbalance and output-space scalability. In multi-label classification, some label combinations may occur rarely, while other label combination overpopulate the label space. This problem affects the model's ability to capture the linguistic pattern, since the labels suffer from an imbalanced distribution. Some attempts (Park et al., 2023; Mao et al., 2024; Zhang et al., 2023) have shown promising results, but we are more interested in the transformation of multi-label classification task into binary classification problem solutions (Jabreel et al., 2019) since it is conceptually simple and highly efficient.

Cross-lingual transfer has become popular since building new, high-quality emotion datasets for low-resource languages is costly and notably difficult. Several works in CLER task (Šmíd et al., 2025; Xiong et al., 2024; Kadiyala, 2024) have demonstrated promising results and have strengthened methodological developments in cross-lingual emotion recognition. Moreover, many successful cross-lingual tasks, particularly those employing knowledge transfer, find that utilizing languages belonging to the same linguistic family significantly enhances performance (Patil et al., 2022; Thakkar et al., 2024).

To date, the paradigm of Cross-Lingual Emotion Recognition has significantly benefited from the advent of LLMs, which move beyond the capabilities of earlier PLMs through massive multilingual pre-training and enhanced capacity for cross-lingual knowledge transfer. Several recent works (Creanga et al., 2025; Xue et al., 2025b; Barfi et al., 2025) have successfully demonstrated the utility of LLMs in solving CLER challenges. While these works demonstrate significant performance gains, the majority of them rely on the capabilities of a single, monolithic LLM.

To optimize the reasoning capabilities of individual language model agents, methods such as Chain-of-Thought (CoT) prompting (Wei et al., 2022) have been introduced to provide a multi-step reasoning process before arriving at a final decision. Furthermore, to address reasoning errors that arise, Pan et al. (2024) proposed a self-correction mechanism to refine its prior response. However, a study (Huang et al., 2024) indicated that a single LLM agent often struggles to effectively correct its response without receiving external feedback.

Zehran et al. (2023) proposed a human-mimicking multi-agent collaboration approach, wherein multiple agents engage in peer discussions, critically analyze each other's outputs, and provide feedback to one another to collectively refine the final solution. The proposed approach obtained a significant result in solving math problems collectively. This shift from isolated self-correction to collaborative peer review motivates our framework, which leverages joint decision-making to improve cross-lingual knowledge transfer for low-resource emotion recognition. We are inspired by the multi-agents collaboration approach proposed in this work. Given its strong performance on mathematical reasoning tasks, we hypothesise that it may also be effective for the CLER task.

Since no Balinese emotion dataset is available, we first build the dataset and annotate its labels manually. This introduces the critical challenge of annotation reliability. Some works attempted to solve this problem by introducing Krippendorff's α (Artstein and Poesio, 2008), Split Half Class Match Percentage (SHCMP) (Mohammad, 2024), and Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013). Krippendorff's α assesses the overall reliability of the annotated dataset rather than the reliability of individual annotators. It is not designed to measure annotator-specific competence. Meanwhile, SHCMP is best suited

for scalar or dimensional annotation schemes, which is unsuitable for our categorical labels. Consequently, we select the MACE model as the most suitable for multi-annotator categorical or multi-label annotation.

3 Methodology

In this section, we introduce the methodological framework developed to address the CLER task in Balinese. We first describe the data creation process such as the specification of the data source, the quality control process, and the application of the MACE model to ensure high annotator reliability. Furthermore, we elaborate on our novel LLM adaptation method. This includes the specific details of the prompt engineering strategy utilizing few-shot examples from linguistically related Austronesian languages, and the architecture of the proposed Multi-Agent Peer Collaboration system.

3.1 Data Source and Pre-processing

We initiate the dataset creation process by collecting textual data from both classic and modern narrative short stories written in Balinese. This source was chosen because narrative fiction provides rich, contextualized textual segments that are highly indicative of emotional states. Prior to the annotation phase, we systematically preprocessed the raw text to ensure data uniformity and reliability. We remove irrelevant punctuation and perform a specific script transformation by converting the character *è* to *e*[3] to ensure uniformity with the target Latin script representation. Furthermore, we removed all text instances containing sensitive information, specifically identifiable person identities, to help ensure unbiased emotion prediction.
Table 1 presents a detailed summary of the corpus statistics, which collectively characterize the properties of the dataset.

3.2 Data Annotation and Annotator Reliability

Three annotators participated in the emotion annotation process, denoted as $A1$, $A2$, and $A3$. Two annotators ($A2$ and $A3$) are Balinese linguistics academics, while one annotator ($A1$) has an academic background in psychology, providing domain knowledge relevant to emotional interpretation. For the emotion annotation task, we utilized INCEpTION (Klie et al., 2018), a web-based platform designed for collaborative text annotation. Annotators were instructed to select all emotion categories that applied to a given text segment. The label set comprised the six basic emotions adopted from Ekman's model: anger, disgust, fear, joy, sadness, and surprise. An additional category, neutral, was assigned if the annotator determined that no basic emotion was present in the text. Table 2 summarizes the distribution of these labels across the dataset.

Statistic	Value
Total Number of Instance(samples)	1260
Total Word Count	15289
Average Sentence Length	12.13
Vocab Size (Unique Words)	3730
Num. of Text with Single Label	703
Num. of Text without Label (Neutral)	390

Table 1: General Corpus Statistics

Emotion Label	Count	Percentage(%)
Anger	176	13.9
Disgust	38	3
Fear	136	10.7
Joy	217	17.2
Sadness	250	19.8
Surprise	229	18.1

Table 2: Label Distribution

We evaluated annotator reliability using the MACE model (Hovy et al., 2013) to ensure the quality of our multi-label emotion annotations. The dataset comprises six emotion categories following Ekman's model, and each instance was manually annotated by three annotators. Because MACE is designed for single-label inputs, we adopted a binary multi-label strategy by running MACE independently for each emotion $c \in \{anger, disgust, fear, joy, sadness, surprise\}$. For each label, MACE estimates an annotator-specific competence score $\gamma_{a,c}$, which reflects the probability that annotator a provides the correct decision for label c. To select a single annotator as the most reliable source of truth, we aggregated competence across all emotion categories for each

[3]Balinese text distinguishes two phonetic forms of the letter *e* : the closed vowel, as in 'bible', written as *é*, and the open vowel, as in 'red', written with the standard *e*.

annotator. Formally, for annotator a, we compute:

$$\Gamma_a = \frac{1}{|C|} \sum_{c \in C} \gamma_{a,c}$$

The annotator with the highest overall competence,

$$a^* = \arg\max_a \Gamma_a$$

was selected as the reference annotator. The final gold labels for each instance were then aligned with the annotation decisions provided by a^*. This approach yields a consistent multi-label dataset grounded in the judgments of the most reliable annotator. In addition to MACE, we compute Krippendorff's alpha (Krippendorff, 2011) to quantify inter-annotator agreement. This metric complements MACE by providing a chance-corrected reliability score that reflects the overall consistency among the annotators.

3.3 Multi-label Classification Transformation to Binary Classification

We reformulate the original multi-class classification problem into a series of independent binary classification tasks, reflecting the complexities outlined in the background section. The method involves creating a separate binary classifier for each emotion label within the predefined label set. For a given text instance, the adaptation of the LLMs is structured to classify whether each specific emotion label exists or does not exist in that text.

This classification is executed by utilizing direct prompt engineering for each emotion $l_i \in L$. Specifically, the LLM agent is provided with the input text and the following structured instruction to facilitate a Chain-of-Thought reasoning path: *"Analyze the text below for the presence of* $\{emotion\}$*. Explain your reasoning briefly and conclude with 'Answer:' followed by either 'yes' or 'no'."*

This structured prompt isolates the prediction of each emotion, requiring the LLM to first justify its reasoning based on the few-shot examples, and then provide a binary conclusion (`'yes'` or `'no'`). Let $L = \{l_1, l_2, \ldots, l_k\}$ be the set of k emotion labels, the model effectively learns k functions, $f_i : X \rightarrow \{0, 1\}$, where $f_i(x) = 1$ indicates the presence of emotion l_i in the text x. This transformation isolates the prediction of each emotion, allowing the multi-agent system to focus its collaborative reasoning on the specific label.

3.4 Few-Shot Cross-Lingual Classification with LLMs

To enable effective transfer of knowledge to the low-resource Balinese language, we employ a few-shot in-context learning strategy that exploits the multilingual capabilities of LLMs. Here, N denotes the total number of demonstrative examples. We strictly require N to be an even integer to ensure a perfectly balanced class distribution within the prompt. Since our task is framed as binary classification for each emotion label (l_i), we provide exactly $N/2$ positive examples (where the emotion is present) and $N/2$ negative examples (where the emotion is absent).

The few-shot examples are sampled from the Austronesian languages contained within the BRIGHTER dataset: Indonesian, Javanese, and Sundanese. We select N examples for each emotion proportionally from these three source languages and enforce class balance by selecting $N/2$ examples that express the target emotion l_i (`'yes'`), and $N/2$ examples that do not express l_i (`'no'`).

3.5 Multi-Agent Peer Collaboration Strategy

We adopted Multi-Agent Peer Collaboration (MAPC) strategy to leverage the heterogeneous strengths of different LLM architectures. We adopt the strategy proposed by Zhenran et al. (2023) to refine classification outcomes collaboratively by introducing three distinct stages:

1. **Initial Solution Creation:** Each independent LLM agent, utilizing the identical few-shot prompt and target Balinese text, creates its initial solution. This solution includes the binary classification (`'yes'/'no'`) regarding the presence of emotion l_i and the Chain-of-Thought reasoning that justifies the prediction based on the in-context examples.

2. **Peer Review and Feedback:** Following the initial solution phase, each agent's full solution is shared with the other peer agents. Each agent then acts as a peer reviewer, critically analyzing the other agents' solutions. This review stage generates constructive feedback that highlights misinterpretations, inconsistencies, or alternative interpretations of the Balinese text's content. Practically, we concatenate each peer solution with the following

prompt: *"Please evaluate the agent's reasoning *step-by-step* and provide direct feedback. Avoid unnecessary text analysis or long explanations—focus strictly on assessing the reasoning"*. For a complete and detailed prompt, see the Appendix 1.

3. **Final Solution Revision:** Given the set of reviews and feedback received from its peers, each agent revises its initial solution. The agent is explicitly instructed to reconsider its original reasoning given other agents' critiques and reviews. We concatenate the reviews with the following prompt: *"Now, using other agents' solutions and feedback below as additional information, Can you conclude whether this text contains the emotion: $\{emotion\}$? Your answer should be Yes/No format. You should also explain your reason for this answer."*

The decision for the emotion l_i is determined by a majority vote across the final revised solutions generated by all participating agents. This consensus mechanism ensures that the final prediction reflects the peer-review process. We formalize the entire process for Cross-lingual emotion recognition on the low-resource Balinese language using the Multi-Agent Peer Collaboration framework. See Algorithm 1 for the detailed procedure of our MAPC framework.

Algorithm 1 Few-Shot Multi-Agent Peer Collaboration for CLER

Input: T: Target Balinese Text instance to be classified
L: Set of Emotion Labels
A: Set of M distinct LLM Agents, $A \in \{a_1, a_2, \ldots, a_M\}$
D: Few-shot Example Dataset
Output: P: Set of Final Predicted Binary Labels for T, $P \in \{0, 1\}$

for each $l \in L$ **do**
  $D_{\text{pos}} \leftarrow \text{RandomPropSample}(D, l)$;
  $D_{\text{neg}} \leftarrow \text{RandomPropSample}(D, \neg l)$;
  $D_l \leftarrow D_{\text{pos}} \cup D_{\text{neg}}$;
  Prompt $\leftarrow$ Concatenate(D_l, T);
  for each $a_i \in A$ **do**
    $S_i \leftarrow \text{AgentInference}(a_i, \text{Prompt})$;
    Initialize Review $R_i \leftarrow \emptyset$;
    for each $a_j \in A, j \neq i$ **do**
      $R_{j,i} \leftarrow \text{AgentReview}(a_j, S_i)$;
      $R_i \leftarrow R_i \cup \{R_{j,i}\}$;
    end for
    $P_{final} \leftarrow \emptyset$;
    $S_{\text{revised},i} \leftarrow \text{AgentRevision}(a_i, S_i, Ri)$;
    $P_{final} \leftarrow P_{final} \cup \{\text{BinaryAnswer}(S_{\text{revised},i})\}$
  end for
  $P_l \leftarrow \text{MajorityVote}(P_{final})$;
  $P \leftarrow P \cup \{P_l\}$;
end for

3.6 Experimental Setup

Evaluation Metrics: To evaluate the performance of the proposed MAPC framework for CLER in the Balinese language, we employ the F-measure ($\mathcal{F}$) as the primary evaluation metric. The F_1-score is particularly suitable for classification tasks involving imbalanced datasets and is calculated as the harmonic mean of precision (P) and recall (R). Since we transformed the native multi-label classification problem into a set of independent binary classification tasks, the standard metrics for evaluating multi-label outputs, such as Jaccard Similarity ($\mathcal{J}$), are not directly applicable. Jaccard similarity is primarily designed to assess the overlap between predicted and true label sets, which is less informative in our decoupled binary setting where performance is assessed on an emotion-by-emotion basis.

Baseline Models: We established several baseline models against which the performance of our MAPC framework is assessed. Our first baseline is the multilingual Pre-trained Language Model (PLM) LaBSE (Language-agnostic BERT Sentence Embeddings). We utilize LaBSE to represent the fine-tuned multilingual models, thereby allowing for a direct comparison between the performance ceiling of traditional PLM architectures and the larger LLMs in the cross-lingual task. Furthermore, we employed a set of single LLM agents as stand-alone baselines to quantify the benefit of the collaboration mechanism. We chose models based on their parameter size and architectures: Kimi K2-Instruct-0905 (with 32 billion active parameters), Qwen 3-32b (with 32 billion parameters), and Llama-3.3-70b (with 70 billion parameters). We evaluate these individual LLM models across zero-shot and few-shot settings.

LLM Inference Settings: During the multi-agent collaboration phase, we employ LLMs as agents. To ensure consistent reasoning while maintaining

adequate diversity in the revision process, we configure the models with the following parameters: temperature = 0.3, maximum tokens = none, and number of few-shot exemplars = 50.

PLM Hyperparameters: The Pre-trained Language Model (PLM) used for initial classification (LaBSE) was fine-tuned using the following hyperparameters: learning rate = $1 \times 10^{-}5$, batch size = 16, epoch = 10, maximum sequence length = 256.

Cross-Lingual Experimental Setting: To address data scarcity in Balinese, we use cross-lingual transfer that leverages its linguistic similarity to related languages. Here, "cross-lingual setting" refers to adapting PLMs and LLMs with examples or training corpora from other Austronesian languages. Specifically, we use Indonesian, Javanese, and Sundanese as source languages to guide the models' representation and interpretation of affect in Balinese.

4 Results and Discussion

In this section, we present the results of the annotator reliability analysis using both the MACE model and Krippendorff's alpha (α). We report annotator competence scores and inter-annotator agreement for each emotion category to examine how annotation reliability varies across emotions in our low-resource emotion dataset, as shown in Table 3. We also demonstrate the comprehensive performance results of our Multi-Agent Peer Collaboration (MAPC) framework in comparison with the established baseline models across the cross-lingual emotion recognition task on the Balinese test set. Table 4 illustrates the average F_1 score computed across all six basic emotion categories of MAPC and all comparison models.

Emotion	**A1**	**A2**	**A3**	$\boldsymbol{\alpha}$
Anger	0.87	0.71	0.72	0.62
Disgust	0.64	0.70	0.55	0.48
Fear	0.76	0.62	0.45	0.45
Joy	0.90	0.80	0.55	0.68
Sadness	0.81	0.67	0.50	0.55
Surprise	0.74	0.68	0.60	0.58
Average	0.78	0.69	0.56	0.56

Table 3: Annotator competence scores ($\gamma_{a,c}$) per emotion using MACE

4.1 Results

Table 3 presents the annotator competence scores estimated using MACE for each emotion category. The results reveal consistent differences in annotator reliability across emotions. Annotator A1 achieves the highest competence scores across all six emotion categories, with particularly strong performance for anger and joy, whereas disgust and fear exhibit comparatively lower reliability across annotators. The emotions joy and anger also display higher inter-annotator agreement (α), which is consistent with the consistently high annotator competence scores estimated by MACE for these categories and likely reflects their relatively clear affective cues. In contrast, disgust and fear show lower agreement, corresponding to greater variability and lower competence estimates among annotators, which is in line with their more subjective interpretation. Based on the averaged competence scores, we select Annotator A1 as the reference annotator for constructing the final labels.

Table 4 summarizes the results of all cross-lingual approaches using Indonesian, Javanese and Sundanese data when target language is Balinese. We observe several key insights into CLER task. Firstly, the traditional multilingual PLM, LaBSE, yielded an average F_1-score of 57.49 confirming its limitation in this task. In contrast, all evaluated LLMs except Kimi K2-Instruct-0905 demonstrated significantly higher zero-shot performance, with the Llama-3.3-70b model achieving the highest zero-shot result at 63.34.

Secondly, the introduction of few-shot examples produced mixed results across the individual LLMs, emphasizing the weakness of simple in-context learning for this specific CLER problem. While Qwen3-32b Few-shot saw a notable increase to 63.81, the performance of Llama-3.3-70b Few-shot surprisingly dropped to 58.29 from its zero-shot baseline of 63.34. A relative decrease of 8.89% was also observed for Kimi K2-Instruct-0905. Moreover, our Multi-Agent Peer Collaboration (MAPC) framework achieved an average F_1-score of 63.95.

4.2 Discussion

The performance demonstrated by the LaBSE multilingual PLM, which only achieved an average F_1-score of 57.49, serves to establish the difficulty of the CLER task on the Balinese language. This result confirms the curse of multilinguality (Con-

Model	Setting	Macro-F_1
Embedding/Traditional baseline		
LaBSE	Zero-shot transfer	57.49
LLM baselines (Qwen3-32b)		
Qwen3-32b (Zero-shot)	Zero-shot	57.99
Qwen3-32b (Few-shot)	Few-shot	63.81
LLM baselines (Llama-3.3-70b)		
Llama-3.3-70b (Zero-shot)	Zero-shot	63.34
Llama-3.3-70b (Few-shot)	Few-shot	58.29
LLM baselines (Kimi K2-instruct-0905)		
Kimi K2-Instruct-0905 (Zero-shot)	Zero-shot	44.53
Kimi K2-Instruct-0905 (Few-shot)	Few-shot	40.57
MAPC (ours)	Few-shot	63.95

Table 4: Performance comparison between the proposed MAPC framework and baseline models across zero-shot and few-shot settings. The table reports average Macro-F1 scores across all six basic emotion categories on the evaluation dataset. Bold values denote the best performance.

neau et al., 2020), a phenomenon where increasing the number of languages in pre-training can lead to a dilution of the model's capacity for any single language, particularly harming low-resource languages whose dedicated parameters are compressed to accommodate high-resource languages. We observe a counter-intuitive pattern in the relationship between model size and cross-lingual performance on the Balinese task when comparing Kimi K2-Instruct-0905, Qwen3-32b and Llama-3.3-70b. In the zero-shot setting, the larger Llama-3.3-70b outperforms Qwen3-32b and Kimi K2-Instruct-0905 achieving an F_1-score of 63.34 compared to 57.99. However, this trend reverses in the few-shot setting, where Qwen3-32b exhibits a substantial performance gain, reaching an F_1-score of 63.81, while Llama-3.3-70b shows a notable decline to 58.29. This opposing performance trajectory suggests that increased model scale does not uniformly translate to improved adaptability in low-resource, few-shot scenarios, and highlights differing sensitivities of large language models to in-context learning signals.

The experimental results indicate that the Multi-Agent Peer Collaboration (MAPC) framework, which yielded an average macro F_1-score of 63.95, consistently attained the highest aggregate performance among all tested LLM baselines in both zero-shot and few-shot settings. To evaluate the reliability of the performance improvements observed in the MAPC framework, we conducted a statistical significance analysis using McNemar's test across each emotional category. The resulting $p-$values were as follows: anger ($p = 0.07$), disgust ($p = 0.13$), fear ($p = 0.24$), joy ($p = 0.47$), sadness ($p = 0.28$) and surprise ($p = 0.34$). Although these values do not meet the conventional threshold of $p < 0.05$ for individual categories, several aspects merit attention. First, the $p-$value for the anger category ($p = 0.07$) indicates a strong directional trend toward significance. This pattern suggests that the MAPC framework may be particularly effective in addressing linguistic complexities in contexts where single-model inference typically exhibits limitations. Second, the MAPC framework appears to integrate the complementary strengths of individual agents, yielding an aggregate output that surpasses the maximum performance attained by any single agent in isolation. Table 5 presents the performance trajectories for each model.

Single Model	Before	After
Qwen3-32b	63.52	67.72
Kimi K2-Instruct-0905	40.13	62.47
Llama-3.3-70b	57.43	61.66

Table 5: Performance comparison between single model in MAPC before and after review and revision process.

The initial Macro-F1 performances of the agents before the collaboration were: Qwen 3 (63.52), Kimi K2-Instruct-0905 (40.13), and Llama-3.3 (57.43). Crucially, the peer review and revision stage corrected these individual failures. This significant improvement in the weakest agent (Kimi K2-Instruct-0905) and the slight boost in the strongest agent (Qwen3-32b) confirm that the peer review process effectively enhances the general rea-

soning quality across the board.

The revision stage of the MAPC framework proved essential in addressing these individual model limitations. Through the structured interaction between agents, initial predictions were re-evaluated, leading to a measurable increase in individual reasoning quality. Despite the lower baseline accuracy of Kimi K2, the MAPC framework achieved a final score of 63.95, surpassing the highest single-model performance of Qwen 3 (63.52 or 63.81 in few-shot). This indicates that the majority voting mechanism, when applied to refined outputs, effectively filters out the stochastic errors present in lower-performing models.

5 Limitations

The use of a majority-voting consensus mechanism in a collaborative framework risks agents converging on a shared error rather than correcting one another. If the majority of agents share a systemic misunderstanding of a specific Balinese cultural idiom, the framework is unlikely to correct the error. The MAPC framework operates by querying multiple LLM agents across several rounds of review and revision. This multi-agent architecture inherently incurs higher computational costs and inference latency compared to single-model inference.

6 Conclusion

In this study, we investigated the application of a Multi-Agent Peer Collaboration (MAPC) framework to address the challenges of cross-lingual emotion recognition in the low-resource Balinese language. A fundamental contribution of this work was the construction of a novel, multi-label Balinese emotion dataset from scratch, which was carefully annotated by reliable experts specializing in Balinese linguistics and psychology. By using this dataset as a target in cross-lingual setting, we confirm that LLMs significantly outperform traditional multilingual PLMs like LaBSE, which are often hindered by the "curse of multilinguality" in such restricted settings. MAPC framework attained the highest aggregate performance ($F_1 = 63.95$) exceeding the results of all single-model baselines in both zero-shot and few-shot configurations. While the margin over the strongest individual model (Qwen3-32B) is modest, the framework's success lies in its ability to consolidate reasoning from diverse architectures into a more cohesive output.

We demonstrated that while individual models may show inconsistent performance or sensitivity to few-shot prompts in low-resource settings, the MAPC review and revision process consistently enhances the reasoning quality of participating agents. By facilitating a multi-stage validation process and a majority-based consensus, the framework effectively filters out individual errors, leading to a more reliable final output. Although statistical significance was not strictly achieved across all emotional categories, the consistent directional improvement suggests that collaborative architectures offer a robust alternative to single-model inference. Future work will focus on further optimizing this collaborative process by exploring intelligent routing mechanisms to dynamically prioritize the most reliable model outputs based on the specific linguistic complexity of the input.

References

AliNadia, TubaishatAbdallah, Al-ObeidatFeras, ShabazMohammad, WaqasMuhammad, HalimZahid, RidaImad, and AnwarSajid. 2023. Towards enhanced identification of emotion from resource-constrained language through a novel multilingual bert approach. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.

Mobin Barfi, Sajjad Mehrpeyma, and Nasser Mozayani. 2025. Unclelm at semeval-2025 task 11: Rag-based few-shot learning and fine-tuned encoders for multilingual emotion detection.

Wei Bi and James T Kwok. 2013. Efficient multi-label classification with many labels.

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 Transactions on Intelligent Systems and Technology*, 15:39.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A.

Adelman, and 2 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications 2025 16:1*, 16:3280–.

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. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Claudiu Creanga, Teodor George Marchitan, and Liviu P. Dinu. 2025. Team unibuc - nlp at semeval-2025 task 11: Few-shot text-based emotion detection.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 11993–12006.

Sabit Hassan, Shaden Shaar, and Kareem Darwish. 2022. Cross-lingual emotion detection. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6948–6958.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Mohammed Jabreel, Antonio Moreno, Mohammed Jabreel, and Antonio Moreno. 2019. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences 2019, Vol. 9,*, 9.

Ram Mohan Rao Kadiyala. 2024. Cross-lingual emotion detection through large language models. *WASSA 2024 - 14th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis, Proceedings of the Workshop*, pages 464–469.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Xuan Liu, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Emotion classification for short texts: an improved multi-label method. *Humanities and Social Sciences Communications 2023 10:1*, 10:306–.

Jian Mao, Kai Huang, and Jinming Liu. 2024. Mlawsmote: Oversampling in imbalanced multi-label classification with missing labels by learning label correlation matrix. *International Journal of Computational Intelligence Systems*, 17:205–.

Saif M. Mohammad. 2024. Worrywords: Norms of anxiety association for over 44k english words. *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 16261–16278.

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, and 29 others. 2025. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. 1:8895–8916.

Muljono, Anisa Sri Winarsih, and Catur Supriyanto. 2016. Evaluation of classification methods for indonesian text emotion detection. *Proceedings - 2016 International Seminar on Application of Technology for Information and Communication, ISEMANTIC 2016*, pages 130–133.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8320–8340.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Wongi Park, Inhyuk Park, Sungeun Kim, and Jongbin Ryu. 2023. Robust asymmetric loss for multi-label long-tailed learning.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:219–233.

Robert Plutchik. 1991. *The emotions*. Bloomsbury Publishing PLC.

Oddy Virgantara Putra, Fathin Muhammad Wasmanson, Triana Harmini, and Shoffin Nahwa Utama. 2020. Sundanese twitter dataset for emotion classification. *CENIM 2020 - Proceeding: International Conference on Computer Engineering, Network, and Intelligent Multimedia 2020*, pages 391–395.

Gaurish Thakkar, Nives Mikelić Preradović, Marko Tadić, Gaurish Thakkar, Nives Mikelić Preradović, and Marko Tadić. 2024. Transferring sentiment cross-lingually within and across same-family languages. *Applied Sciences 2024, Vol. 14,*, 14.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. *NewSumm 2023 - Proceedings of the 4th New Frontiers in Summarization Workshop, Proceedings of EMNLP Workshop*, pages 12–23.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Feng Xiong, Jun Wang, Geng Tu, and Ruifeng Xu. 2024. Hitsz-hlt at wassa-2024 shared task 2: Language-agnostic multi-task learning for explainability of cross-lingual emotion detection. *WASSA 2024 - 14th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis, Proceedings of the Workshop*, pages 476–482.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023. Towards reasoning in large language models via multi-agent peer review collaboration. *Preprint*, arXiv:2311.08152.

Jieying Xue, Phuong Minh Nguyen, Minh Le Nguyen, and Xin Liu. 2025a. JNLP at SemEval-2025 Task 11: Cross-Lingual Multi-Label Emotion Detection Using Generative Models. pages 20–27.

Jieying Xue, Phuong Minh Nguyen, Minh Le Nguyen, and Xin Liu. 2025b. Jnlp at semeval-2025 task 11: Cross-lingual multi-label emotion detection using generative models.

Daniel Yohanes, Jessen Surya Putra, Kenneth Filbert, Kristien Margi Suryaningrum, and Hanis Amalia Saputri. 2023. Emotion detection in textual data using deep learning. *Procedia Computer Science*, 227:464–473.

Ruohong Zhang, Yau Shian Wang, Yiming Yang, Donghan Yu, Tom Vu, and Likun Lei. 2023. Long-tailed extreme multi-label text classification by the retrieval of generated pseudo label descriptions. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*, pages 1092–1106.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:11189–11204.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025. Advancing cross-lingual aspect-based sentiment analysis with llms and constrained decoding for sequence-to-sequence models. *International Conference on Agents and Artificial Intelligence*, 2:757–766.

A Appendix

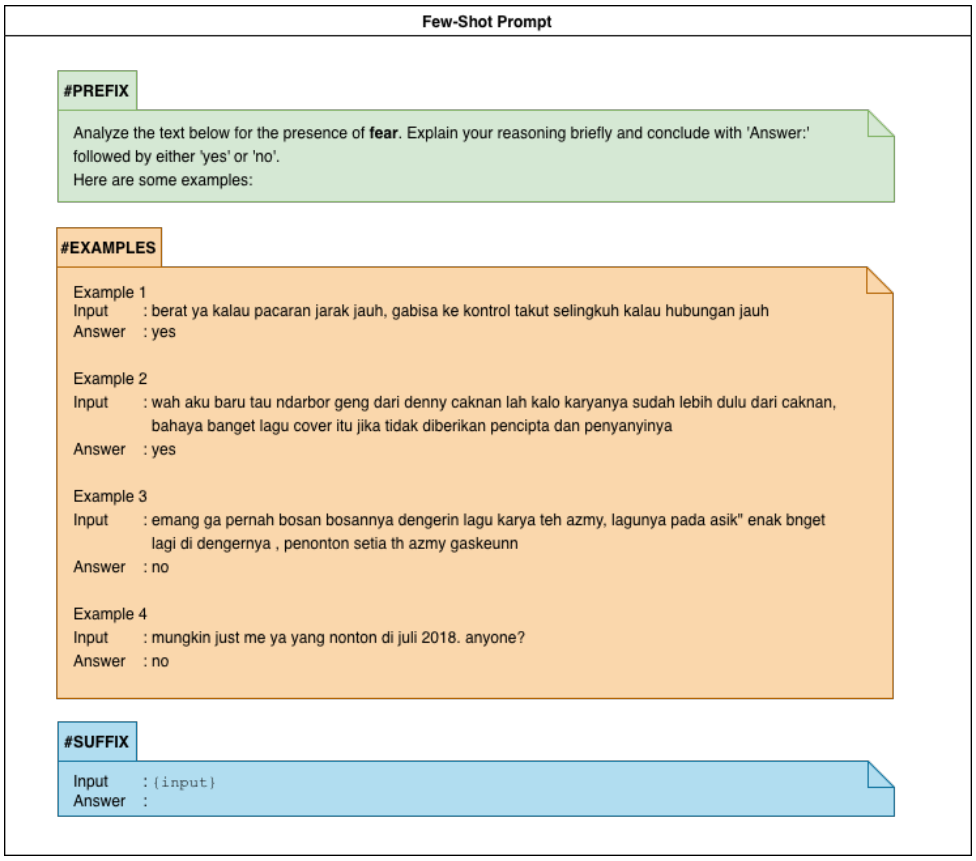

Figure 1: Few-shot prompt example of assessing emotion fear.

Review Prompt

#PREFIX

You are reviewing the reasoning text of an Emotion Recognition Agent. Your task is to provide concise, structured, and objective feedback on the agent's reasoning quality. The agent should detect emotion by using the provided examples only as an initial reference, and by carefully examining the contextual cues in the text being analyzed.

#EXAMPLES

Example 1
Input : berat ya kalau pacaran jarak jauh, gabisa ke kontrol takut selingkuh kalau hubungan jauh
Answer : yes
Conclusion : The emotion: **fear** is present in this text

Example 2
Input : wah aku baru tau ndarbor geng dari denny caknan lah kalo karyanya sudah lebih dulu dari caknan, bahaya banget lagu cover itu jika tidak diberikan pencipta dan penyanyinya
Answer : yes
Conclusion : The emotion: **fear** is present in this text

Example 3
Input : emang ga pernah bosan bosannya dengerin lagu karya teh azmy, lagunya pada asik" enak bnget lagi di dengernya , penonton setia th azmy gaskeunn
Answer : no
Conclusion : The emotion: **fear** is present in this text

Example 4
Input : mungkin just me ya yang nonton di juli 2018. anyone?
Answer : no
Conclusion : The emotion: **fear** is present in this text

#SUFFIX

Please evaluate the agent's reasoning *step by step* and provide direct feedback.
Avoid unnecessary text analysis or long explanations, focus strictly on assessing the reasoning.

Original text analyzed: `{text}`
Target emotion: **fear**
Agent's perception about the presence of fear in this text: `{answer}`
Agent's reasoning: `{reason}`

Return your final output in exactly this structure and nothing else:
My Review = {{"feedback": <string>, "confidence": <string>}}

Notes:
- "feedback" should be a concise critique (3–5 sentences) highlighting correctness, gaps, or faulty logic.
- "confidence" is your confidence level on your feedback. It must be from 1–10 and in string format.

Figure 2: Review prompt example of assesing emotion fear.

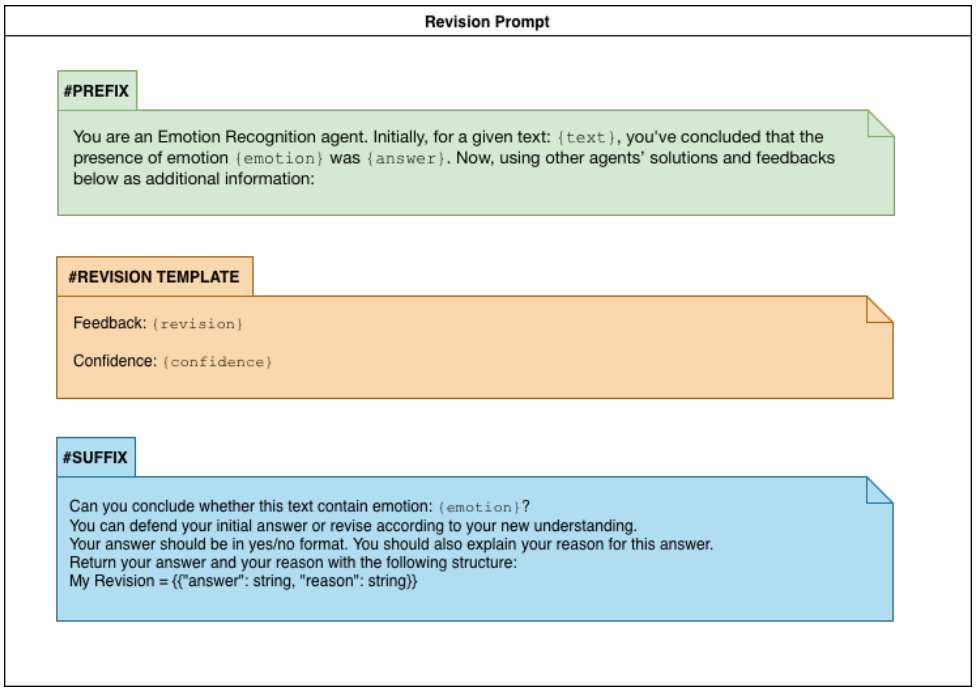

Figure 3: Revision prompt example of assesing emotion fear.