# FeedForward at SemEval-2024 Task 10: Trigger and sentext-height enriched emotion analysis in multi-party conversations

**Zuhair Hasan Shaik**[1], **R Dhivya Prasanna**[1], **Enduri Jahnavi**[1],
**Rishi Koushik Reddy Thippireddy**[1], **P S S Vamsi Madhav**[1],
**Sunil Saumya**[1], and **Shankar Biradar**[1]
[1]Department of Data Science and Intelligent Systems,
Indian Institute of Information Technology Dharwad, Dharwad, Karnatka, India
(zuhashaik12, prasanna0083,jahnavienduri,rishikoushik18,pssmvamsi)
@gmail.com (sunil.saumya, shankar)@iiitdwd.ac.in

## Abstract

This paper reports on an innovative approach to Emotion Recognition in Conversation and Emotion Flip Reasoning for the SemEval-2024 competition with a specific focus on analyzing Hindi-English code-mixed language. By integrating Large Language Models (LLMs) with Instruction-based Fine-tuning and Quantized Low-Rank Adaptation (QLoRA), this study introduces innovative techniques like Sentext-height and advanced prompting strategies to navigate the intricacies of emotional analysis in code-mixed conversational data. The results of the proposed work effectively demonstrate its ability to overcome label bias and the complexities of code-mixed languages. Our team achieved ranks of 5, 3, and 3 in tasks 1, 2, and 3 respectively. This study contributes valuable insights and methods for enhancing emotion recognition models, underscoring the importance of continuous research in this field.

## 1 Introduction

Emotional analysis has come quite a long way. In the context of natural language processing (NLP), history reveals an evolution of the emotion analysis task. The task has always been about recognizing emotions from text, evolving from those early-day systems that were able to recognize emotions from standalone text (Akhtar et al., 2019; Chatterjee et al., 2019; Mageed and Ungar, 2017; Shankar Biradar and Chauhan, 2021) to the current cutting-edge challenge of Emotion Recognition in Conversation (ERC) (Lei et al., 2023; Hazarika et al., 2018). Well-designed simple methods have demonstrated that recognizing the emotion of a user's expression enables a broad range of practical applications in diverse domains, from e-commerce (Gupta et al., 2013) to healthcare (Khanpour and Caragea, 2018).

ERC plays a significant role in illustrating how the emotion change during the interpersonal communications. By contrast to the isolation of single texts, ERC struggles with how emotions shift through a combination of different speakers in conversation. Motivated by the urgent need to understand the complex interactions of emotions during dialogue, a new issue has arisen—Emotion-Flip Reasoning (EFR) (Kumar et al., 2022a, 2024b). EFR is a novel Endeavour aiming at identifying precisely which utterances transform an emotion within a person's flow of speech. Apart from just emotions, EFR seeks to unravel the complexities of emotion flips, offering valuable insights into the dynamics of human interaction. Emotional flips can result from internal party interactions or from external elements such as speaker gestures or verbal messages.

The practical importance of EFR extends beyond theoretical limitations. In reality, it has applications in a variety of sectors. EFR plays a crucial part in the development of reward and punishment systems, as well as interpretable emotion recognition systems. Further, the widespread use of Hindi-English code-mixed language online shows the cultural change. NLP is facing new challenges in the accurate identification of emotions in a dynamic cultural context. Language switching during the conversation makes the work of emotion recognition systems even more complex. Further building an adaptable system capable of capturing the subtle variations in emotions that emerge in such a hybrid language setting is the need of the hour.

In order to promote research in this field, the organisers of SemEval 2024, Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) [1] organised a shared task. The organisers created three sub-tasks:

- Task 1: Emotion Recognition in Conversation (ERC) in Hindi-English code-mixed conversations

[1] https://lcs2.in/SemEval2024-EDiReF/

*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 745–756
June 20-21, 2024 ©2024 Association for Computational Linguistics

- Task 2: Emotion Flip Reasoning (EFR) in Hindi-English code-mixed conversations

- Task 3: EFR in English conversations.

The following is an illustration of the definition for the ERC and EFR tasks:

- Emotion Recognition in Conversation (ERC) is focused on assigning emotions to individual utterances or phrases within a dialogue. It involves analyzing conversation data to identify the emotional states expressed by speakers throughout the interaction. The goal of ERC is to accurately recognize and categorize the emotions conveyed in each utterance.

- Emotion Flip Reasoning (EFR) aims to identify triggers for emotion flips in multi-party conversations. A trigger can be caused by one or more utterances, and some emotion flips might not be triggered by other speakers but by the target utterance itself (self-trigger). EFR analyzes dialogue data to understand the causes behind shifts in emotions, providing insights into the dynamics of emotional exchanges in conversations.

Our team, FeedForward, participated in Semeval-2024 task 10 and achieved rankings of 5, 3, and 3 in subtasks 1, 2, and 3, respectively[2]. For detailed insights and findings regarding this task, please refer to the task description paper of SemEval-2024 Task 10 (Kumar et al., 2024a).To tackle this problem, we propose state-of-the-art techniques such as Sentext-height for emotion recognition in multi-party conversations and ratio-wise splitting in trigger datasets for the EFR task. Additionally, we utilized instruction-based QLoRA training of 7-billion-parameter models for both ERC and EFR tasks.

The outline of the article is as follows: Section 2 offers an in-depth exploration of the background study. In addition, Section 4 comprehensively discusses the proposed methodologies. Finally, the experimental outcomes are illustrated in section 5.

## 2 Related work

Emotion detection in the standalone text is a well-known challenge in the Natural Language Processing domain (Akhtar et al., 2019; Chatterjee et al., 2019). However, unlike single text, emotion recognition in conversation data requires numerous complicated understandings of contextual information and speakers (Wagh and Sutar, 2023). In accordance with this, the majority of studies used deep neural networks with memory functions to solve sophisticated understandings of conversational text data (Hazarika et al., 2018; Weston et al., 2014). Furthermore, the developers of (Zhong et al., 2019) attempt to include the role of speakers into the conversational model by using memory networks during two-party discussions.

The utilization of external information is also vital in recognizing emotions in multi-party conversations. The authors of (Wen et al., 2023) proposed the DIMMN network for capturing speaker interaction information during multi-party conversations, in addition to text, audio, and video aspects during experiments. Conventional categorical label-based approaches fail to capture quantitative measurements of emotion; to solve this issue, the authors of (Yang et al., 2023) created a low-dimensional cluster-level contrastive learning model incorporating linguistic and factual information. Furthermore, the (Li et al., 2023) established a discourse link between utterances by adding symbolic information into multi-party interactions.

ERC in low-resource code-mixed text has received little attention. The authors of (Ghosh et al., 2023; Saumya et al., 2022) created a Hindi-English emotion-annotated corpus and established a transformer-based end-to-end framework with multitask learning. Furthermore, most existing studies only account for emotion recognition, but very few studies looked beyond emotion recognition to interpret the results. In one such study, (Kumar et al., 2022b; Fharook et al., 2022), the authors introduced a novel Emotion-Flip Reasoning (EFR), which aims to identify past utterances that have triggered one's emotional state to flip at a certain time, in addition to ERC.

## 3 Dataset

### 3.1 MaSac_ERC

The organizers of EDiReF of SemEval 2024 have provided the MaSac_ERC dataset (Kumar et al., 2023) for emotion recognition in Hindi-English Code-Mixed Conversations (Task 1). The task is to recognize emotions for speaker utterances in conversations. The train dataset contains 343 conversations and a total of 8506 utterances, which

[2] All proposed models are openly available at: https://huggingface.co/collections/zuhashaik/multi-party-dialoz-65d34c9f74e0888ef4e66da3

contain 8 emotion classes—Neutral, Joy, Anger, Sadness, Contempt, Fear, Surprise, and Disgust. The data set is significantly skewed, and the distribution of emotions across the train, validation, and test data is shown in Figure 1.

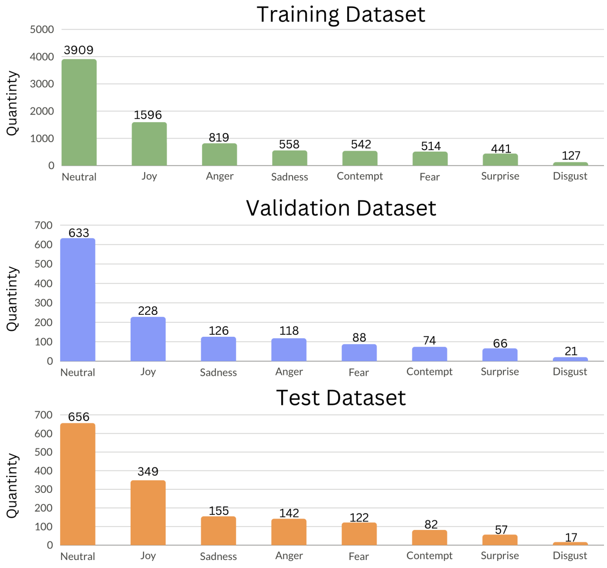

Figure 1: Emotion distribution of the MaSac_ERC

### 3.2 MaSac_EFR and MELD_EFR

The organizers of EDiReF of SemEval 2024 have provided the MaSac_EFR and MELD_EFR datasets Emotion Flip Reasoning (EFR) in Hindi-English code-mixed conversations (Task 2) and English Conversations (Task 3) respectively. The goal is to find all utterances that trigger a flip in the emotion of a speaker within a conversation. The MaSac_EFR train dataset contains 4,893 conversations having 6,542 triggers and 92,233 non-triggers. And the dataset distribution is clearly illustrated in Table 1. Similarly the MELD_EFR dataset contains 4,000 conversations having 5,575 triggers and 29,425 non-triggers. And the data distribution of triggers and non-triggers is illustrated in Table 2.

| Trigger | Train | Validation | Test |
|---|---|---|---|
| Yes (1) | 6542 | 434 | 416 |
| No (0) | 92233 | 7024 | 7274 |

Table 1: MaSac_EFR Label distribution

| Trigger | Train | Validation | Test |
|---|---|---|---|
| Yes (1) | 5575 | 494 | 1169 |
| No (0) | 29425 | 3028 | 7473 |

Table 2: MELD_EFR Label distribution

## 4 Methodology

In this section, a comprehensive study of the methodology employed, focusing on Emotion Recognition in Conversations (ERC) in Hindi-English code-mixed data and Emotion Flip Reasoning (EFR) in Hindi-English code-mixed conversations, as well as in English for Task1, Task2, and Task3, respectively, for SemEval-2024 Shared Task-10.

### 4.1 Task 1 : ERC in Hinglish

In this study, the focus lies on examining emotions within Hindi-English (Hinglish) code-mixed multi-party conversations using advanced language models. Various methods are explored, including refining BERT derivatives and translating code-mixed utterances for emotion classification. Furthermore, strategies like simplifying emotion labels and utilizing large language models with effective prompts are implemented to improve performance.

#### 4.1.1 BERT derivatives as Baseline

As is commonly known, BERT (Devlin et al., 2019) demonstrates exceptional proficiency in sentiment analysis across various domains in natural language processing (NLP). However, the dataset comprises Hindi-English code-mixed text, necessitating pre-trained BERT derivatives capable of understanding Hinglish.

After an extensive exploration and experimentation phase with various BERT models, several BERT derivatives trained on Hindi or Hindi-English code-mixed datasets were identified. These include bert-base-multilingual-cased[3]**??**, l3cube-pune's hing-mbert-mixed-v2 (Joshi, 2023), lxyuan's distilbert-base-multilingual-cased-sentiments-student, and papluca's xlm-roberta-base-language-detection. Additionally, google's FNet-base (Lee-Thorp et al., 2022) was considered due to its substantial research presence in sentiment analysis, showcasing promising outcomes.

In this approach, each utterance paired with its corresponding emotion was treated as a data point extracted from the MaSac_ERC dataset. Subsequently, this data was utilized to fine-tune BERT derivatives for the emotion classification task, irrespective of its position within the conversation sequence and relevant contextual nuances.

[3] https://huggingface.co/google-bert/bert-base-multilingual-cased

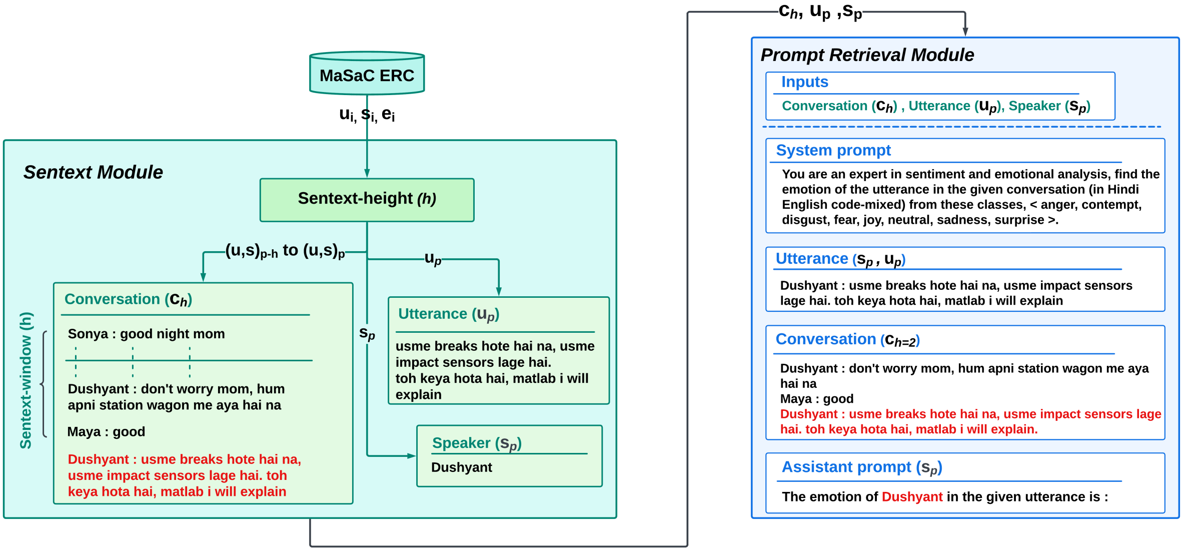
MaSaC ERC
$u_i$, $s_i$, $e_i$
Sentext Module
Sentext-height (h)
$(u,s)_{p-h}$ to $(u,s)_p$
$u_p$
$s_p$
Sentext-window (h)
Conversation ($c_h$)
Sonya : good night mom
Dushyant : don't worry mom, hum apni station wagon me aya hai na
Maya : good
Dushyant : usme breaks hote hai na, usme impact sensors lage hai. toh keya hota hai, matlab i will explain
Utterance ($u_p$)
usme breaks hote hai na, usme impact sensors lage hai. toh keya hota hai, matlab i will explain
Speaker ($s_p$)
Dushyant
$c_h$, $u_p$ ,$s_p$
Prompt Retrieval Module
Inputs
Conversation ($c_h$) , Utterance ($u_p$), Speaker ($s_p$)
System prompt
You are an expert in sentiment and emotional analysis, find the emotion of the utterance in the given conversation (in Hindi English code-mixed) from these classes, < anger, contempt, disgust, fear, joy, neutral, sadness, surprise >.
Utterance ($s_p$ , $u_p$)
Dushyant : usme breaks hote hai na, usme impact sensors lage hai. toh keya hota hai, matlab i will explain
Conversation ($c_{h=2}$)
Dushyant : don't worry mom, hum apni station wagon me aya hai na
Maya : good
Dushyant : usme breaks hote hai na, usme impact sensors lage hai. toh keya hota hai, matlab i will explain.
Assistant prompt ($s_p$)
The emotion of Dushyant in the given utterance is :

Figure 2: In the overview of the SP Module, the figure illustrates the complete process from slicing conversations with the Sentext Module to obtaining a training-ready prompt from the Prompt Retrieval Module.

In this approach, all layers of the models were retained unfrozen, converging into a feedforward network and subsequently a dense 8-way classifier, empowered by softmax.

### 4.1.2 Hinglish to English Translation

In the study focusing on Emotion Recognition in Code-Mixed Hindi-English Conversations (ERC), a unique methodology was employed. Rather than following a sequential conversation analysis, the code-mixed utterances were transliterated and then translated using IndicXlit (Madhani et al., 2023) and IndicTrans2 (Gala et al., 2023), respectively from AI4BHARAT organization. The inference of the models and the procedure of converting Hinglish to English are accessible here.[4] The translated utterance with its corresponding emotion was then used as a data point to fine-tune BERT and FNet for the sequence classification task.

### 4.1.3 Split and concat

In the split and concat approach, the label was coarse-grained (Neutrals, Negatives, Positives) to study the nuances created by the labels and the dataset complexity. Then, Fine grained to only Negatives (Anger, Sadness, Contempt, Fear, and Disgust) and only Positives (Joy, Surprise) were considered.

The main aim of this approach is to create a ensemble architecture (a classifier's tree) that will reduce the complexity of the dataset for the models being used. At the first level, it classify sentences as Neutral, Negative, or Positive. Then, at the second level, it further classify negatives and positives.

For instance, at the first level, An *NNP* (Neu-Neg-Pos) classifier predicts the sentiment of the utterance as Neutral, Negative, or Positive. If it's Neutral, the process stops there as we already classifed the emotion. Otherwise, it proceeds to the corresponding output sentiment classifier (Negs or Pos) to further classify the fine grained emotion.

### 4.1.4 7Bs enhanced with SP-module

When traditional approaches failed to yield satisfactory results, primarily due to label bias and the complexity of the Hindi-English code-mixed language, which struggled to distinguish between classes effectively, the focus shifted to large language models. 7-Billion (7B) parameter Large Language Models (LLMs) were utilized, taking these models from the shelf and then finetuning using Quantized Low Rank Adaptation QLoRA (Dettmers et al., 2023) on the dataset with effective prompts.

**Sentext-height**

To enhance the model's performance, a novel concept called *Sentext-height* was introduced. Sentext-height is a new idea that comes from context related to sentiment analysis within a sentence. It determines how many previous

[4] The proposed methodology can be found here: https://github.com/Zuhashaik/Multi-Party-DialoZ

utterance influence the emotion analysis of the present utterance in a given conversation. With this, it is possible to capture the emotion state of a speaker in the past utterances, which can contribute to finding the emotion of the present speaker's utterance.

### Prompt-engineerning

LLMs have proven to be significantly reliable for a wide array of tasks in the domain of NLP. While they show significant promise, effective usage requires a carefully curated input. Through extensive experimentation with prompt structures on the foundational models, a conclusion was reached with a prompt that effectively works for the model.

The structure of the Prompt Retrieval Module:

- System prompt: Defines the LLMs role and expected behavior within the interaction, guiding its response.
  <|*system*|>*You are an expert in sentiment and emotional analysis, find the emotion of the utterance in the given conversation (in Hindi-English code mixed) from these classes, [anger, contempt, disgust, fear, joy, neutral, sadness, surprise].*

- Utterance: This contains the present utterance ($u_p$) with the respective speaker ($s_p$) attached to it before the utterance.
  <|*utterance*|> *{Speaker}:{Present_utterance}*

- Conversation: This has the conversation that is driven by sentext-height ($h$). It consists of $h$+1 utterances with Sentext-window ($u_{p-h}$ to $u_{p-1}$) along with the current utterance $u_p$ which to be evaluated, each with their corresponding speakers identified to indicate who made those utterances.
  <|*conversation*|> *{conversation, h}*

- Assistant prompt: Provides an incomplete statement or scenario and expects LLM to finish the very next word, making it a classification task that we're interested in.
  <|*assistant*|>*The emotion of {Speaker ($s_p$)} in the given utterance is :*

  In this case, the probable choices are the various emotions listed in the system prompt. These emotions include anger, contempt, disgust, fear, joy, neutral, sadness, and surprise. The model tries to classify within these emotion categories.

Data preprocessing hence concludes with the setting the sentext-height and selection of the appropriate prompt, collectively referred to as the SP-module (Sentext-Prompt) and clearly illustrated in the Figure 2.

### QLoRA and Instruction Finetuning

After preparing the data with the SP-module, we used the prompt-processed dataset to fine-tune 7Bs with Instruction-based QLoRA for classifying emotions. We made 6 datasets, altering the sentext-height (h) from 2 to 7. Each model will train on every dataset, and we'll choose the best sentext-height based on how well the model performs. The models employed in this proposed study include Llama-2-7b-chat-hf (Touvron et al., 2023), zephyr-7b-beta (Tunstall et al., 2023), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), and openchat_3.5 (Wang et al., 2023).

Due to the challenges posed by *Catastrophic forgetting* (Luo et al., 2023) and computational constraints, the full training of LLMs (7Bs) cannot be carried out. Instead, QLoRA was chosen. This method involves quantizing the model during inference and then applying LoRA. With LoRA, the model parameters are frozen, and an additional low-rank matrix is introduced beside the attention layer weights, rather than training all parameters. This approach significantly reduces training time and memory requirements, often resulting in improved performance compared to traditional fine-tuning methods.

Additionally, a custom classifier was designed, where the last decoder layer in the 7B LLM is connected to an 8-way dense network powered by a softmax classifier. This is distinct from the text-generation LLM, where the 7B LLM is connected to a vocab-sized (32,000 in this case) classifier to predict the next word of the given input, which iterates until the end of sequence tag <eos> arises or the token limit is reached.

The total integration of the Sentext-height, Prompt-module and custom architecture with LoRA are demostrated in the figure 3.

### Experimental setup

In this case, all models are inferred and trained in FP16 (Half-precision, float16). Following extensive experiments with various sentext-height

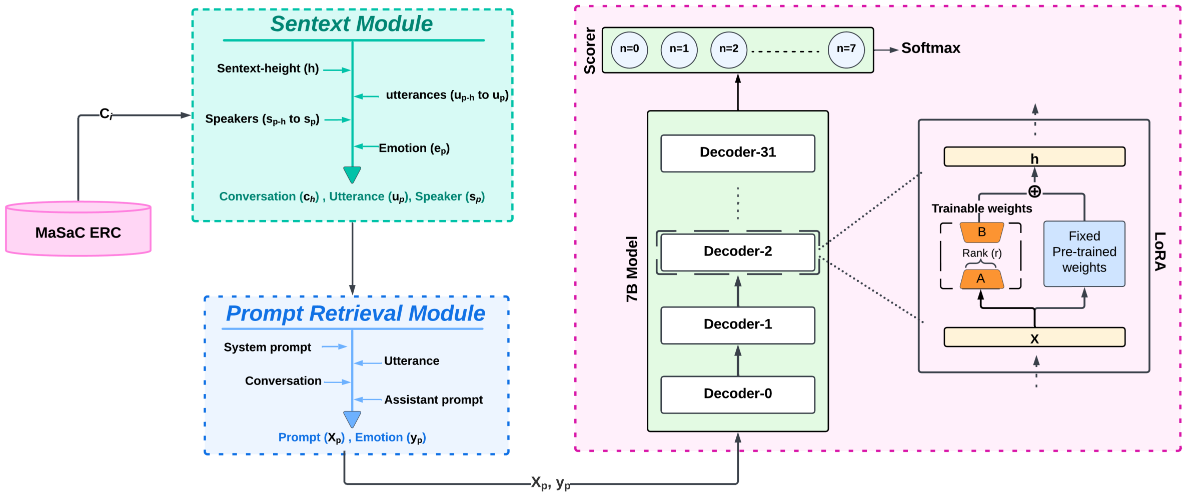

Figure 3: The MaSac-ERC-Z framework, figure displays how the Sentext Module and Prompt Retrieval Module are combined with a 7-billion parameter LLM. It also shows how LoRA is incorporated into the model, with each decoder having a low-rank matrix next to the pre-trained attention weights. This LoRA technique is applied specifically to all 32 decoder layers.

(h={2-7}), the hyperparameters that proved effective for the proposed model have been identified, as outlined in Table 3. Considerable RAM and computing capabilities are leveraged, supported by 3×32G Nvidia Tesla V100 GPUs.

| Hyper parameter | Value |
|---|---|
| Rank (LoRA config) | 16 |
| LoRA Alpha (LoRA config) | 64 |
| Dropout (LoRA config) | 0.2 |
| Learning Rate | $2 \times 10^{-5}$ |
| Learning Rate Scheduler | Constant |
| Batch size | 1 |
| Gradient acumulation step | 1 |
| adam_beta1 | 0.9 |
| adam_beta2 | 0.999 |
| adam_epsilon | $1.000 \times 10^{-8}$ |
| rms_norm_eps | $1.000 \times 10^{-5}$ |

Table 3: Hyper parameters for Training 7Bs

## 4.2 Task 2 and 3 : EFR in Hinglish and English respectively

Since both task 2 and 3 involve Emotion Flip Reasoning but in different languages, maintaining the core model while adjusting the input is proposed. When providing embeddings to the model, rich semantic information from the text in the same language as the dataset is ensured. This approach enables obtaining language-aware contextual embeddings for the core model under development.

### 4.2.1 Attention-Based Utterance Fusion

In this approach, the Bert-based embeddings $(e_1, e_2, ..e_n)$ are extracted for each and every utterance $(u_1, u_2, ..u_n)$ in the conversation of n utterances. Consider $u_p$ as the present utterance from the conversation, and the task is to determine whether it is the trigger for the $u_n$ utterance which led to an emotion flip. Now, $u_p$ and $u_n$ are considered, and their embeddings $e_p$ and $e_n$ respectively are obtained. These embeddings are then linearly concatenated and passed through multi-head attention to capture intricate patterns within concatenated utterance pairs. Subsequently, a feed-forward network followed by a binary classifier is applied. Experimentation has been conducted with different BERT derivatives and the number of heads in multi-head attention has been varied.

### 4.2.2 7Bs for EFR

Following the MaSac-ERC-Z framework used in Task 1 with the 7B language model, the similar architecture is adopted here. However, a 2-way dense softmax classifier is incorporated instead of 8, as the task aims for binary classification (trigger or non-trigger). Furthermore, the focus is solely on identifying triggers rather than analyzing conversational emotion, so the sentext module is omitted. Additionally, a specialized prompt module is introduced to enhance the efficiency of trigger retrieval for the specific task.

**Prompt Module**

After extensive experimentation in the playground of the foundational models, a prompt that works effectively for this task was concluded.

The glance of the prompt:

- System prompt: Defines the LLMs role and expected behavior within the interaction, guiding its response.
  <|*system*|>*In your role as an expert in sentiment and emotion analysis, your primary objective is to identify trigger utterances for emotion-flips in multi-party conversations (in Hindi-English code-mixed). Evaluate the provided dialogue by analyzing changes in emotions expressed by speakers through their utterances. Your task is to determine the accuracy of the hypothesis based on these emotional shifts.*
  For Task 3, which is the MELD dataset in English, *(in Hindi-English code-mixed)* from the system prompt is removed, and the remaining architecture will remain the same.
- Hypothesis: This contains the hypothesis and expecting the LLM to evaluate the hypothesis.
  <|*Hypothesis*|> *The utterance <{present_utterance}> is a trigger for the emotion-flip in <{speaker}'s> : <{final_utterance}> in the conversation*
- Conversation: This section contains the entire conversation, ensuring no chance of missing context. The emotions are also provided immediately after each utterance in the conversation, which is crucial for identifying the emotion flip and analyzing which utterance is the trigger.
  <|*conversation*|> *{conversation}, {emotions}*
- Assistant prompt: A sentence is left incomplete, assuming that the LLM has already generated something related to the input task. The expectation is for the LLM to complete this sentence.
  <|*assistant*|> *The given Hypothesis is :*

**Instruction and QLoRA finetuning**

As discussed, this approach follows a similar method proposed in Task-1, the MaSac-ERC-Z module, where a dataset is created using the prompt module and Instruction-based QLoRA fine-tuning is performed for the 7B model on the dataset. However, the constraint is that there are 98,775 (6,542 Triggers and 92,233 Non-triggers) and 35,000 (5,575 Triggers and 29,425 Non-triggers) datapoints from the MaSac_EFR and MELD_EFR datasets respectively. This can significantly slow down the trainings and take a lot of time to complete. Experimentations would be impractical under these circumstances. To avoid these constraints, the dataset was sliced into an 1:n ratio of Triggers to Non-Triggers, where n = {1,2,..}.

For instance, in the Task 2 dataset (MaSac_EFR), there are 6,542 triggers and 92,233 non-triggers. To preserve all triggers, the same number of non-triggers was selected to create a 1:1 dataset, yielding 13,084 datapoints from a total of 98,775. Similarly, for a 1:2 ratio, 6,542 triggers were retained, and 13,084 non-triggers were selected, and so forth up to a 1:3 ratio. This reduction in dataset size resulted in shorter training times leading to more efficient model training.

## 5 Results

This section presents a comprehensive study on outcomes from Task 1, Task 2, and Task 3. The weighted-f1 score is used as the standard metric for all tasks, as recommended by the task organizers and utilized to evaluate the submission hosted on Codalab.

### 5.1 Task 1

**The baseline**

In the proposed study, BERT derivatives were utilized, among which mBERT exhibited significant performance, yielding a weighted F1 score of 41.70. Consequently, this served as an initial baseline for evaluating the effectiveness of subsequent ideas and models. The corresponding scores are provided in Table 4.

| **Base-Model** | **Weighted-F1** |
|---|---|
| mBERT | **41.70** |
| hing-mbert-mixed-v2 | 28.76 |
| lxyuan | 40.25 |
| papluca | 37.39 |
| fnet-base | 38.08 |

Table 4: Weighted-F1 scores of Finetuned models

**Translation**

Following the initial efforts, the aim was to enhance performance further, considering the intricate nature of deciphering patterns within code-mixed languages. The approach involved converting Hinglish (a mix of Hindi and English) into English and using transformer-based models to identify the emotions. After this transformation, a weighted-f1 score of 40.03 was achieved with bert-base-uncased and 35.79 with fnet-base. The decline is assumed to be the accumulation of errors across three key processes: transliteration, translation, and classification. These processes inherently carry a high risk of errors, which likely impacted the classification accuracy.

### The classifier tree

In the proposed work, *Split and concat* in Task1, the impact of coarse and fine-grained approaches on classification was analyzed. This examination aimed to pinpoint areas for improvement in achieving scores above the baseline. The primary challenge lies in classifying Neutrals within the complex Hindi-English code-mixed context, resulting in a weighted-f1 score of 55.16. Additionally, categorizing fine-grained negatives poses a significant challenge, as evidenced by a weighted-f1 score of 39.87. However, identifying positives proves comparatively easier, with weighted-f1 of 91.28.

The strengths of all three classifiers were combined, resulting in an aggregate score of 41.46 with BERT-Tree. The Ensemble BERT-Tree consists of Hing-BERT as the first-level classifier (NNP) and mBERT for further classifying negatives (Negs) and positives (Pos) at the second level. These models were chosen based on their performance scores in both coarse and fine-grained classification tasks. Nonetheless, this represents a decline from the baseline as discussed earlier. The decrease may be due to the compounding errors from each classifier that affect the final classification.

The detailed investigation of the study is outlined in Table 5. In the table, "Neutral-Negative-Positive" represents the coarse grain classification, while "Negatives (Negs)" and "Positives (Pos)" indicate the fine grain emotion categories.

### 7Bs enhanced with SP-module

In the analysis, the proposed approaches fell short of delivering satisfactory results, preventing the achievement of a weighted-f1 score in the 50s. The complexity of code-mixed languages posed a significant challenge, and the methods struggled to grasp the nuances of context and sentiments effectively.

| **Base model** | **NNP** | **Negs** | **Pos** |
|---|---|---|---|
| mBERT | 49.42 | **39.87** | **91.28** |
| hing-mbert | **55.16** | 11.80 | 79.47 |
| lxyuan | 49.71 | 39.01 | 90.69 |
| papluca | 53.80 | 32.89 | 90.35 |
| fnet-base | 48.69 | 26.70 | 89.91 |
| **W-F1 (ALL)** | | | |
| BERT-Tree (En) | | **41.46** | |

Table 5: Weighted-F1 scores of Coarse and Fine-Grained Emotion Classification Results, combined all to construct a tree like classifer to classify all 8 emotions. 'En' denotes Ensemble here.

However, upon transitioning to 7Bs for this task and conducting extensive experimentation on various foundational models and sentext-height (choosing n between 2-7), a threshold of 50s was finally surpassed, which elevated the system to the 5th position in the competition. Specifically, a weighted-f1 score of 51.17 was attained using the Zephyr-7b-beta model with a sentext-height of 3.

Based on the analysis presented, Table 6 and figure 4 illustrates the performance of various 7B models across different sentext-height (h) values.

| **7B Models** | **Sentext-height (h)** | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 |
| llama2 | 49.0 | **49.5** | 48.3 | 49.0 | 48.3 | 47.9 |
| zephyr | 46.7 | **51.2** | 45.5 | 46.0 | 47.4 | 46.3 |
| mistral | 45.5 | 45.5 | 44.5 | **46.1** | 45.5 | 47.0 |
| openchat | 42.4 | 46.7 | 47.3 | 45.0 | 43.2 | **48.6** |

Table 6: Weighted-F1 scores of 7B models with different Sentext-height (h) values.

### All-Together

Bringing everything together for Task 1, the study began with mBERT as the benchmark, followed by efforts to refine performance through translation and ensemble techniques, which encountered challenges and resulted in reduced scores. Further exploration into classification strategies revealed difficulties in nuanced identification, leading to mixed outcomes. Finally, incorporating SP-modules helped to surpass the 50s score threshold, reflecting progress in addressing the complexities of code-mixed languages. The comprehensive results for Task 1 can be viewed in Table 7, providing a detailed overview of the study's findings.

| Model | Model Names | W-F1 | Method |
|---|---|---|---|
| Encoder-Only | mBERT | 41.7 | Seq-cls |
| | hing-mbert | 28.76 | Seq-cls |
| | lxyuan | 40.25 | Seq-cls |
| | papluca | 37.39 | Seq-cls |
| | fnet-base | 38.08 | Seq-cls |
| | BERT | 40.03 | Translation |
| | FNet | 35.79 | Translation |
| | BERT-Tree | 41.46 | Ensemble |
| Decoder-Only | llama_h3 | 49.52 | QLoRA |
| | mistral_h3 | 45.5 | QLoRA |
| | **zephyr_h3** | **51.17** | **QLoRA** |
| | openchat_h3 | 46.73 | QLoRA |
| | mistral_h5 | 46.07 | QLoRA |
| | openchat_h7 | 48.58 | QLoRA |

Table 7: The table provides weighted F1 scores comparison of various methods and models. For Decoder-only models, the sentext-height is specified after the model's name. "Seq-cls" denotes sequence classification.

### 5.1.1 Task 2 and Task 3

**7Bs for EFR**

In Task 1, 7Bs demonstrated remarkable performance, motivating the extension of their use to EFR. As outlined in the Methodology, training requires a significant amount of time due to the large number of data points. To address this, the dataset was sliced and implemented a 1:n ratio (Triggers : Non-Triggers), resulting in (1+n)x datapoints (where x represents the number of triggers).

This concept was applied to Task 3 as well, given the similar nature of Task 2 but with English-language data, ensuring consistency in the approach across both tasks.

**Task2**

| **7B models** | **1:1** | **1:2** | **1:3** |
|---|---|---|---|
| openchat | 57.81 | 55.60 | 58.51 |
| zephyr | 66.19 | 66.32 | 76.96 |

Table 8: Task 2, Weighted-F1 scores of 7B models with different splitting ratios

**Task3**

| **7B models** | **1:1** | **1:2** | **1:3** |
|---|---|---|---|
| openchat | 71.52 | 71.29 | 72.53 |
| zephyr | 70.77 | 71.97 | 71.91 |

Table 9: Task 3, Weighted-F1 scores of 7B models with different splitting ratios

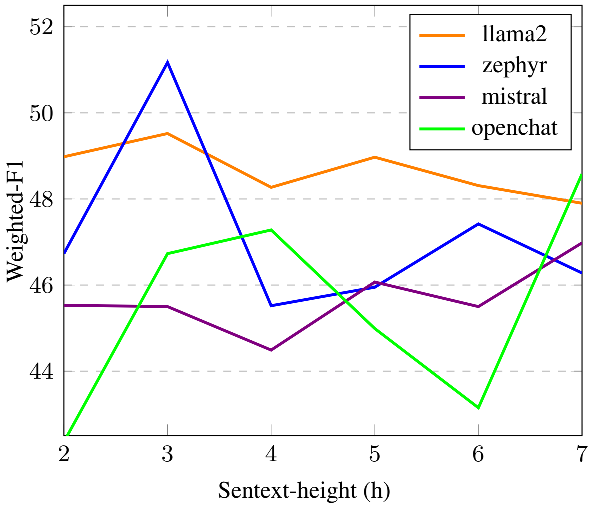

Figure 4: Graphical representation of the performance of 7B models with different Sentext-height (h) values.

For task 3, we considered the validation set and trained with a specific ratio (3:1) and model (openchat_3.5) that resulted in the highest score (72.53), achieving a weighted F1 score of 73.94 From the demonstrated experiments, the ratio of 1:3 yielded the highest scores in both Task 2 and Task 3, resulting in securing the 3rd rank in both tasks respectively. The results of the experiments with various ratio's and 7B models is given in the table 8 for MaSac_EFR which is task2 and table 9 for MELD_EFR which is task3.

## 6 Conclusion

This paper discusses the proposed work for the competition EDiReF SemEval-2024 hosted on Codalab. The study mainly focuses on emotion and emotion flip-trigger analysis specifically within multi-party conversational data. Through innovative approaches and the utilization of state-of-the-art techniques such as Large Language Models (LLMs), Instruction-based fine-tuning, and Quantized Low-Rank Adaptation (QLoRA), our team achieved promising results in Emotion Recognition in Conversation (ERC) and Emotion Flip Reasoning (EFR) tasks. However, obstacles persist, especially in addressing label bias and capturing nuanced emotions in Hindi-English code-mixed language. The findings underscore the need for further research to enhance model performance, ultimately improving emotional analysis in conversational data.

## References

Md Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2019. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE transactions on affective computing*, 13(1):285–297.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Shaik Fharook, Syed Sufyan Ahmed, Gurram Rithika, Sumith Sai Budde, Sunil Saumya, and Shankar Biradar. 2022. Are you a hero or a villain? a semantic role labelling approach for detecting harmful memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 19–23.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Multitasking of sentiment detection and emotion recognition in code-mixed hinglish data. *Knowledge-Based Systems*, 260:110182.

Narendra Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio. 2013. Emotion detection in email customer care. *Computational Intelligence*, 29(3):489–505.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

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

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024a. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref). In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Shivani Kumar, Shubham Dudeja, Md Shad Akhtar, and Tanmoy Chakraborty. 2024b. Emotion flip reasoning in multiparty conversations. *IEEE Transactions on Artificial Intelligence*, 5(3):1339–1348.

Shivani Kumar, Ramaneswaran S, Md Akhtar, and Tanmoy Chakraborty. 2023. From multilingual complexity to emotional clarity: Leveraging commonsense to unveil emotions in code-mixed dialogues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9638–9652, Singapore. Association for Computational Linguistics.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022a. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022b. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2022. Fnet: Mixing tokens with fourier transforms.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework.

Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul NC, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023.

Aksharantar: Open indic-language transliteration datasets and models for the next billion users.

Muhammad Abdul Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Sunil Saumya, Vanshita Jha, and Shankar Biradar. 2022. Sentiment and homophobia detection on youtube using ensemble machine learning techniques. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR.*

Sunil Saumya Shankar Biradar and Arun Chauhan. 2021. mbert based model for identification of offensive content in south indian languages. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR.*

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Nanda R Wagh and Sanjay R Sutar. 2023. Enhanced emotion recognition for women and children safety prediction using deep network. *International Journal of Intelligent Systems and Applications in Engineering*, 11(10s):500–515.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data.

Jintao Wen, Dazhi Jiang, Geng Tu, Cheng Liu, and Erik Cambria. 2023. Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion*, 91:123–133.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.

# A Performance of Top model for Task 1

In the following appendix, we present the performance metrics of the instruction-tuned zephyr-7b-beta model (zephyr_h3) with Sentext-height (h=3) which is the top performing model with a Weighted-F1 of 51.17 in sub task 1, Emotion Recognition in Conversation (ERC).

## A.1 Confusion Matrix

The confusion matrix in figure 5 visually represents the performance of the zephyr_h3 model in classifying different emotions. We observed a notable amount of confusion primarily between the emotions of joy and neutral, which could be attributed to the prevalence of neutral expressions in the dataset. This suggests a bias towards categorizing ambiguous or mild emotions as neutral, potentially impacting the accuracy of our predictions.

Additionally, there appears to be confusion between the emotions of anger and fear, as well as between contempt and sadness. These overlaps indicate potential similarities in the facial expressions or textual cues associated with these emotions, highlighting areas where our model may require further refinement.

## A.2 Classification Report

The comprehensive classification report in table 10 for the zephyr_h3 model, showcasing precision, recall, F1 score, and support across various emotions.

The report further underscores the performance of our model across different emotions. While achieving relatively high precision for joy and neutral emotions, indicating a good ability to correctly identify these categories, our model struggles with emotions such as disgust and fear, as evidenced by lower precision scores. This indicates a tendency

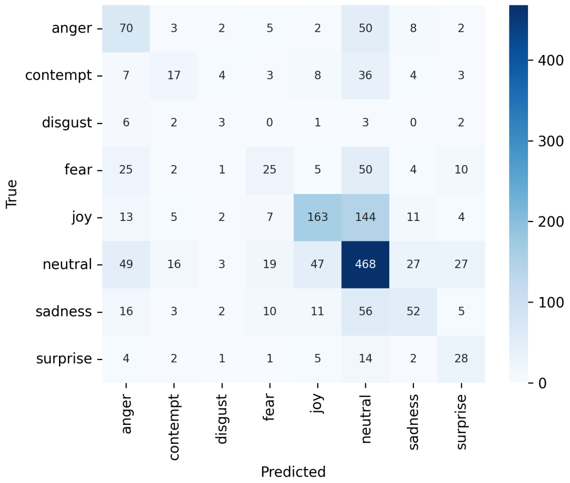

Figure 5: Confusion Matrix

for the model to misclassify instances of these emotions as other classes. Moreover, the overall accuracy of our model is moderate, indicating room for improvement in effectively distinguishing between the diverse emotional states. These findings emphasize the importance of addressing biases in the dataset and further fine-tuning the model to enhance its ability to accurately classify a wider range of emotions.

| **Emotion** | **Precision** | **Recall** | **F1 Score** | **Support** |
|---|---|---|---|---|
| Anger | 0.38 | 0.52 | 0.44 | 142 |
| Contempt | 0.33 | 0.20 | 0.25 | 82 |
| Disgust | 0.19 | 0.18 | 0.18 | 17 |
| Fear | 0.31 | 0.20 | 0.24 | 122 |
| Joy | 0.69 | 0.48 | 0.56 | 349 |
| Neutral | 0.58 | 0.72 | 0.65 | 656 |
| Sadness | 0.47 | 0.35 | 0.40 | 155 |
| Surprise | 0.34 | 0.47 | 0.39 | 57 |
| **Accuracy** | | 0.53 | | |
| **Macro Avg** | 0.41 | 0.39 | 0.39 | 1580 |
| **Weighted-Avg** | 0.53 | 0.53 | 0.52 | 1580 |

Table 10: Classification Report

### A.3 Performance summary

Our classification model demonstrates moderate overall accuracy of 53%, with strengths in identifying joy and neutral emotions, boasting precision scores of 69% and 58% respectively. However, it struggles with emotions such as disgust and fear, showing lower precision scores of 19% and 31% respectively. Confusion primarily arises between joy and neutral emotions, possibly due to dataset biases towards neutral expressions. Further refinement is needed to enhance the model's ability to accurately classify a broader spectrum of emotions. The macro-average F1 score is 39%, while the weighted average F1 score is 52%, indicating room for improvement in capturing the nuances of different emotional states.