UCSC NLP at SemEval-2024 Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)

Steven Au, Decker Krogh, Esha Ubale, and Neng Wan
University of California, Santa Cruz
Baskin School of Engineering, Natural Language Processing
{sttau, dkrogh, eubale, newan}@ucsc.edu

Abstract

We describe SemEval-2024 Task 10: EDiReF consisting of three sub-tasks involving emotion in conversation across Hinglish code-mixed and English datasets. Subtasks include classification of speaker emotion in multiparty conversations (Emotion Recognition in Conversation) and reasoning around shifts in speaker emotion state (Emotion Flip Reasoning). We deployed a BERT model for emotion recognition and two GRU-based models for emotion flip reasoning[1]. Our model achieved F1 scores of 0.45, 0.79, and 0.68 for subtasks 1, 2, and 3, respectively.

1 Introduction

Emotion recognition in natural language provides quantifiable insights into the traditionally qualitative realm of emotive language, bridging fields such as psychology, cognition, and linguistics. The explosion of textual data in recent years from social media platforms like Twitter and the introduction of highly cable text-processing models has provided researchers the opportunity to perform analyses on conversations that are highly complex. Despite these developments, the inherent subjectivity of emotion continues to present a challenge to the field.

Without visual information and speech audio, NLP systems must decipher rapid changes in emotional states solely through text, missing out on the nuanced non-verbal cues that often signal these shifts during spoken interactions. The absence of these cues can lead to model inaccuracies during conversational transitions such as from joy to sarcasm, or from calmness to anger.

Understanding the dynamics of emotion in the context of conversations is vital for building better conversational agents. While classifying changes in the emotion of a speaker is an important first step in this goal, it comes up short of being able to explain why the change occurred. Emotion flip reasoning is a task which has been proposed which seeks to identify the specific cause of speaker emotion flips in the context of a conversation (Kumar et al., 2022) (Kumar et al., 2024b). For example if a speaker's emotion in one utterance is joy but in their next utterance it is sad, we would like to pinpoint which utterances in the conversation caused it whether it be another speaker's or their own.

1.1 Hinglish

Hinglish, a blend of Hindi and English written in the Roman alphabet, incorporates English words into traditional Hindi contexts. This code-mixing phenomenon is becoming increasingly prevalent as English extends its influence into non-English speaking societies. Hinglish provides a challenge to models that have only been trained on English and Hindi because the model struggles to distinguish between English and Hindi words (Solorio et al., 2014). Recent work has sought to improve model performance on code-mixed dialog and new datasets have been constructed to enable these developments (Kumar et al., 2023). One goal of this research is to contribute to this work of producing models that can better understand the emotion of speakers in Hinglish.

Commonsense discernment between languages plays a pivotal role in emotion recognition within code-mixed languages, as it aids in navigating the nuanced linguistic landscapes that arise when languages intertwine (Kumar et al., 2023). Historically, individual words and phrases have been identified as significant emotional triggers, serving as fundamental elements in the computational understanding of emotions (Mohammad and Turney, 2010). This is especially pertinent in code-mixed contexts where the semantic layers are compounded by the interplay of distinct linguistic systems.

A large number research on emotion recognition

[1] https://github.com/deckerkrogh/semeval-2024-10

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1492–1497

to date has focused on extracting and interpreting common emotion-laden lexicon from Twitter corpora. While beneficial, this approach predominantly captures public, social media-expressed sentiments, which may not fully encapsulate the subtleties found in personal or private conversational contexts. This is particularly true for code-mixed interactions, where cultural contexts and language mixing patterns can greatly affect emotional expression. Datasets produced from television shows allow for insight into these more these private conversational contexts.

2 Task Description

The **EDiReF shared task at SemEval 2024** is an amalgamation of three subtasks tasks-

(i) Emotion Recognition in Conversation (ERC) in Hindi-English code-mixed conversations,

(ii) Emotion Flip Reasoning (EFR) in Hindi-English code-mixed conversations, and

(iii) EFR in English conversations.

ERC Definition: Given a dialogue, ERC aims to assign an emotion to each utterance from a predefined set of possible emotions.

EFR Definition: Given a dialogue, EFR aims to identify the trigger utterance(s) for an emotion flip in a multi-party conversation dialogue.

Speaker	Utterance	Emotion	Trigger
Sp1	Aaj to bhot awful day tha!		
	(I had an awful day today!)	Sad	0
Sp2	Oh no! Kya hua?		
	(Oh no! What happened?)	Sad	0
Sp1	Kisi ne mera sandwich kha liya!		
	(Somebody ate my sandwich!)	Sad	0
Sp2	Me abhi tumhare liye new bana deti hun!		
	(I can make you a new one right now!)	Joy	1
Sp1	Wo great hoga! Thanks!		
	(That would be great! Thanks!)	Joy	0

Table 1: Example of a dialogue from the MaSaC dataset.

We are one of 84 teams which submitted an entry to the task 10 code submission, and one of 21 which submitted papers for the task 10 workshop. (Kumar et al., 2024a).

2.1 Datasets

Two datasets were used in this task.

MELD is a dataset released in 2017 made from dialog from the TV show *Friends*. This dataset contains a list of conversations, each with multiple utterances that have been tagged with an emotion label. It has seen extensive use in research related to emotion recognition and its use in the finetuning of transformers has produced models with major improvements in tasks such as emotion recognition. For this task the task organizers produced a modified MELD dataset which has been labeled with emotion triggers (Kumar et al., 2024b). This was used for task three.

MaSaC is a Hinglish dataset produced in 2021 containing conversations with emotion-labelled utterances which were extracted from the television show *Sarabhai vs. Sarabhai*. (Bedi et al., 2021). MaSaC was used for tasks one and two. The task organizers tagged emotion triggers for the dataset used in task two.

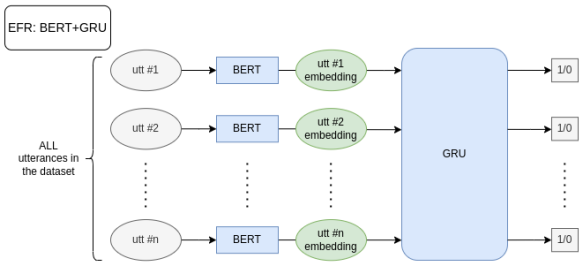

Figure 1: Illustration of the BERT+GRU architecture for the Emotion Flip Reasoning task

3 System Overview

In this study, we introduce an integrated framework that combines Bidirectional Encoder Representations from Transformers (BERT) with Gated Recurrent Unit (GRU) networks.

3.1 BERT for Emotion Recognition in Conversation

BERT was used to perform emotion recognition for the ERC task. Unlike traditional models that process text sequentially, BERT examines text bidirectionally, allowing for a comprehensive understanding of word semantics in context. Additionally, BERT's pre-training on extensive language corpora equips it with a broad understanding of language nuances, idioms, and the varied syntax used to express emotions, providing a robust starting point for fine-tuning emotion-specific datasets.

3.2 BERT+GRU for Emotion Flip Reasoning

Upon extracting contextual embeddings from BERT, we employ a GRU layer to analyze the sequence of conversational utterances. GRUs are a type of recurrent neural network optimized for

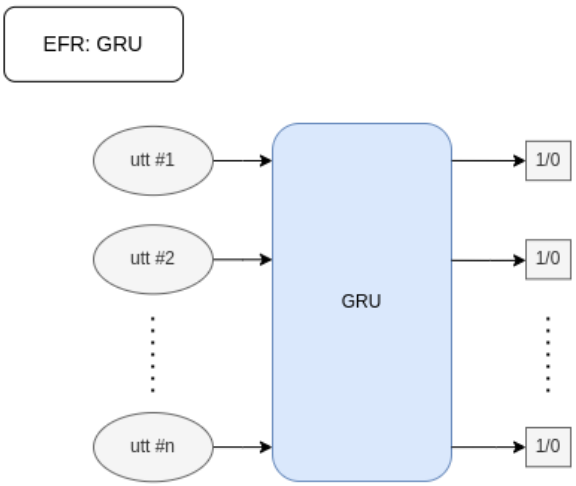

Figure 2: Simple GRU model for Emotion Flip Reasoning.

handling sequential information while mitigating issues related to long-term dependency recognition. The intent is that GRUs will be able to track the evolution of emotional states across a dialogue. Understanding the temporal sequence and the transition between emotional states is necessary in the context of emotion flip detection.

3.3 Rationale for Architectural Integration

The decision to integrate BERT with GRU stems from a strategic consideration of their respective strengths in handling different aspects of emotion analysis. BERT's contextual embeddings provide a snapshot of the emotional landscape within each utterance. While BERT excels at static context understanding, it cannot provide the sequence-to-sequence operation necessary to perform trigger classification. The purpose of the GRU is to use these static BERT embeddings to interpret the flow and dynamics of emotions through time so in order to perform trigger classification.

3.4 GRU

In addition to the BERT+GRU architecture, we also created a simpler GRU model which takes the utterances directly as input.

4 Experimental Setup

4.1 Emotion Recognition Model: BERT

We employ the pretrained BertForSequenceClassification model from the Hugging Face Transformers library. We added a fully connected linear layer with an output that matches the number of emotions. There are 7 emotions for MELD and 8 for MaSaC. The model was finetuned for 4 epochs with the AdamW optimizer set to a learning rate of 5^{-4} and an epsilon of 1^{-5}.

4.2 EFR: BERT+GRU

We constructed a deep learning model utilizing the Keras framework tailored for binary classification tasks.

BERT Embeddings: We first generate embeddings for each utterance in a conversation which will then be fed into the GRU. These embeddings were generated with the same pretrained BERT model used in the ERC task. The goal is that these embeddings can capture and provide emotion-specific information for the GRU in trigger classification.

GRU: The model consists of two bidirectional GRU layers with 32 units. We did not perform any separation between the conversations. Conversational structure is collapsed into a long sequence of utterances and passed into the GRU.

Classifier Layer: The final layer is a dense classifier with a single output unit which performs binary trigger classification for each utterance.

4.3 EFR: GRU

This model is the same as the GRU+BERT model, however instead of using BERT embeddings we use a default Keras embedding layer and pass utterances in directly.

5 Results

Table 2: F1 Scores and Task Placement

Task 1	Task 2	Task 3
0.45 (8)	0.79 (2)	0.68 (8)

5.1 Sub Task-1: ERC in Hindi-English Code-mixed Conversations

The model obtained an F1 score of 0.45 on emotion recognition in Hindi-English code-mixed conversations. The model showed a mediocre ability to capture emotional expressions.

5.2 Sub Task-2: EFR in Hindi-English Code-mixed Conversations

The GRU-only model demonstrated strong performance, achieving an F1 score of 0.76 on the validation Set and 0.79 on the test Set. This was significantly higher than the BERT+GRU model which achieved an F1 score of 0.66. These results suggest that the BERT embeddings were not able provide useful context for the GRU. It also suggests that the GRU is capable of effectively capturing the dynamic nature of emotional transitions. Despite our more novel model performing worse than the simpler one, the simple GRU achieved second place in the CodaLab competition.

5.3 Sub Task-3: EFR in English Conversations

The GRU-only model achieves F1 scores of 0.68 for the validation and 0.67 on the test set on the EFR task, outperforming the BERT+GRU model.

5.4 Further Testing

We tested additional inputs where we passed the speaker information to see if emotion recognition improved in subtasks 2 and 3 for the task. No improvements were seen in F1 but might influence the embeddings. We didn't test the EFR pipeline unless we saw improvement in ERC. We also double-checked the abnormally high emotion recognition F1 score for subtask 2 as we stripped the conversation structure and passed in duplicate utterances with shuffling. We redid the test with unique utterances and achieved .87 F1 in the test set. Surprisingly this did not affect our score for subtasks one or three. We also increased the dataset size by combining datasets 1 and 2 for Hinglish and using the whole dyadic conversations from MELD with dataset 3. The scores for ERC show no changes.

Further testing is necessary to investigate why the GRU-only model outperformed BERT+GRU. We hypothesize that it may be that the GRU simply wasn't large enough to be capable of using the the large BERT embeddings.

Another change to the model that may improve performance is to create some sort of separator embedding between the conversations. This extra information may improve performance by allowing the model to learn where triggers are placed relative to the start and end of a conversation.

6 Conclusion

The baseline approach, which employs basic embeddings of utterances and emotions, proved adequate for capturing emotion flip reasoning in large datasets, despite cultural differences and code-mixing ambiguities. The need to pass conversational information is not a substantial indicator of the prediction EFR triggers nor does passing the emotion labels into the embeddings.

In the future, we plan to explore multitask classification with BERT to determine if combined training enhances the transformer's ability to learn emotional sequence information or integrate additional conversational context to account for speaker dependencies, similar to EmoBERTa's approach. We may attempt to replicate EmoBerta's methodology to see how much emotion labels increase EFR accuracy.

Addressing the challenges of code-mixing in Hinglish and enhancing cross-cultural emotion comprehension remain critical for improving the recognition of emotional transitions.

References

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *arXiv.org*.

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. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-based systems*, 240:108112–.

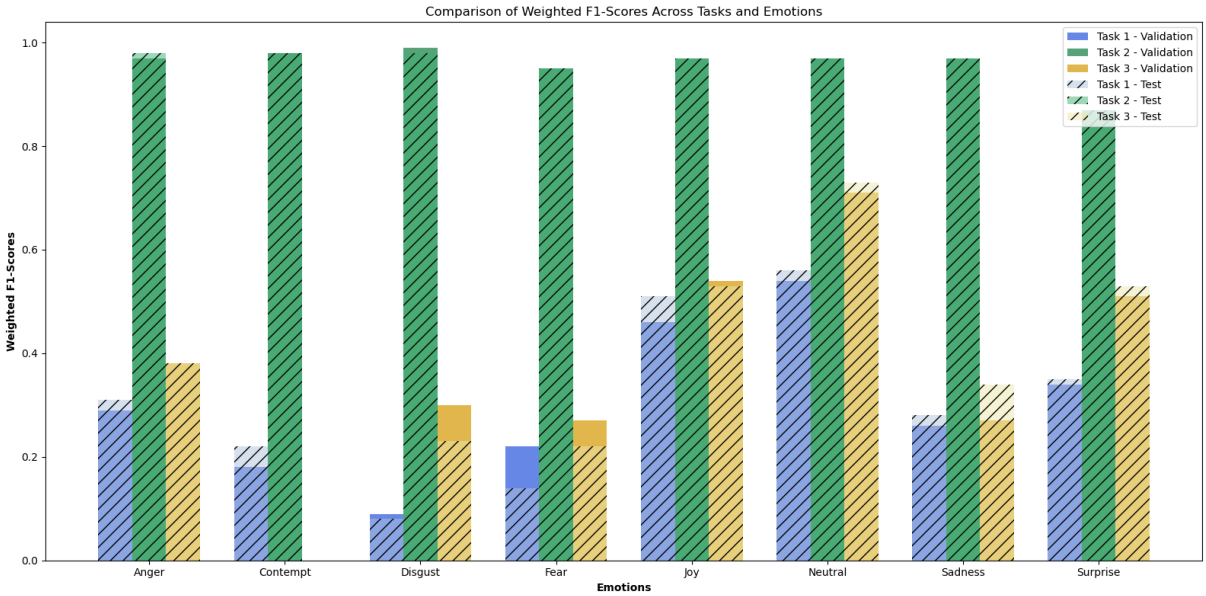

Figure 3: Comparing weighted F1 for emotions ERC across subtasks.

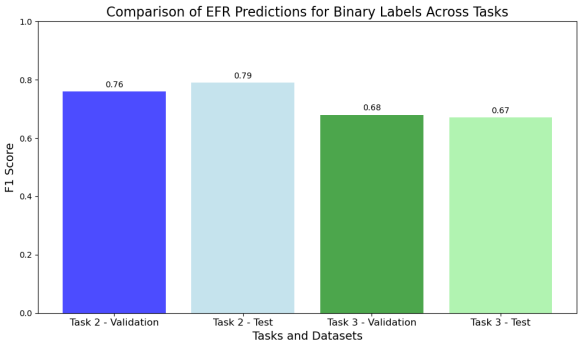

Figure 4: Blue = Subtask 2, Green = Subtask 3

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

A Task Performance Metrics

Tables 4 - 13 show tables of the performance metrics across each task for the validation and test set.

Table 3: Performance Metrics for Task 1 - ERC on Validation Set

Emotion	Precision	Recall	F1-Score	Support
Anger	0.28	0.31	0.29	118
Contempt	0.22	0.15	0.18	74
Disgust	1.00	0.05	0.09	21
Fear	0.29	0.17	0.22	88
Joy	0.45	0.47	0.46	228
Neutral	0.49	0.60	0.54	633
Sadness	0.30	0.23	0.26	126
Surprise	0.31	0.39	0.34	66
Accuracy		0.48		
Macro Avg	0.43	0.30	0.31	1354
Weighted Avg	0.47	0.48	0.46	1354

Table 4: Performance Metrics for Task 1 - ERC on Test Set

Emotion	Precision	Recall	F1-Score	Support
Anger	0.31	0.31	0.31	142
Contempt	0.25	0.20	0.22	82
Disgust	0.14	0.06	0.08	17
Fear	0.18	0.11	0.14	122
Joy	0.52	0.50	0.51	349
Neutral	0.52	0.60	0.56	656
Sadness	0.33	0.25	0.28	155
Surprise	0.29	0.46	0.35	57
Accuracy		0.45		
Macro Avg	0.32	0.31	0.31	1580
Weighted Avg	0.43	0.45	0.44	1580

Table 5: Performance Metrics for Task 2 - ERC on Validation Set

Emotion	Precision	Recall	F1-Score	Support
Anger	0.98	0.97	0.97	639
Contempt	0.99	0.98	0.98	493
Disgust	0.99	0.99	0.99	87
Fear	0.98	0.92	0.95	478
Joy	0.98	0.97	0.97	1801
Neutral	0.96	0.98	0.97	3159
Sadness	0.96	0.97	0.97	487
Surprise	0.90	0.84	0.87	318
Accuracy			0.97	
Macro Avg	0.97	0.95	0.96	7462
Weighted Avg	0.97	0.97	0.97	7462

Table 6: Performance Metrics for Task 2 - ERC on Test Set

Emotion	Precision	Recall	F1-Score	Support
Anger	0.98	0.97	0.98	749
Contempt	0.99	0.98	0.98	547
Disgust	0.99	0.97	0.98	70
Fear	0.96	0.93	0.95	445
Joy	0.97	0.96	0.97	1730
Neutral	0.96	0.98	0.97	3265
Sadness	0.97	0.97	0.97	536
Surprise	0.90	0.84	0.87	348
Accuracy			0.96	
Macro Avg	0.97	0.95	0.96	7690
Weighted Avg	0.96	0.96	0.96	7690

Table 7: Performance Metrics for Task 3 - ERC on Validation Set

Emotion	Precision	Recall	F1-Score	Support
Anger	0.44	0.33	0.38	482
Disgust	0.33	0.28	0.30	64
Fear	0.26	0.28	0.27	156
Joy	0.53	0.56	0.54	597
Neutral	0.66	0.76	0.71	1360
Sadness	0.34	0.23	0.27	343
Surprise	0.51	0.51	0.51	520
Accuracy			0.55	
Macro Avg	0.44	0.42	0.43	3522
Weighted Avg	0.53	0.55	0.54	3522

Table 8: Performance Metrics for Task 3 - ERC on Test Set

Emotion	Precision	Recall	F1-Score	Support
Anger	0.46	0.32	0.38	1215
Disgust	0.36	0.17	0.23	305
Fear	0.18	0.29	0.22	177
Joy	0.50	0.56	0.53	1376
Neutral	0.71	0.76	0.73	3784
Sadness	0.36	0.32	0.34	712
Surprise	0.52	0.54	0.53	1073
Accuracy			0.57	
Macro Avg	0.44	0.42	0.42	8642
Weighted Avg	0.56	0.57	0.56	8642

Table 9: Performance Metrics for Task 2 - EFR on Validation Set

Class	Precision	Recall	F1-Score	Support
False	0.98	0.99	0.99	7028
True	0.81	0.72	0.76	434
Accuracy			0.97	
Macro Avg	0.90	0.86	0.87	7462
Weighted Avg	0.97	0.97	0.97	7462

Table 10: Performance Metrics for Task 2 - EFR on Test Set

Class	Precision	Recall	F1-Score	Support
False	0.99	0.99	0.99	7274
True	0.82	0.76	0.79	416
Accuracy			0.98	
Macro Avg	0.90	0.88	0.89	7690
Weighted Avg	0.98	0.98	0.98	7690

Table 11: Performance Metrics for Task 3 - EFR on Validation Set

Class	Precision	Recall	F1-Score	Support
False	0.94	0.96	0.95	3028
True	0.71	0.66	0.68	494
Accuracy			0.91	
Macro Avg	0.83	0.81	0.82	3522
Weighted Avg	0.91	0.91	0.91	3522

Table 12: Performance Metrics for Task 3 - EFR on Test Set

Class	Precision	Recall	F1-Score	Support
False	0.95	0.96	0.95	7473
True	0.71	0.64	0.67	1169
Accuracy			0.92	
Macro Avg	0.83	0.80	0.81	8642
Weighted Avg	0.91	0.92	0.91	8642