SemEval 2024 – Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)

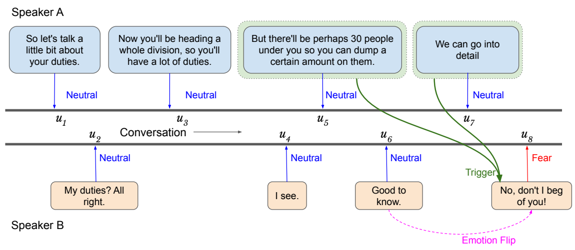
Shivani Kumar[1], **Md Shad Akhtar**[1], **Erik Cambria**[2], **Tanmoy Chakraborty**[3]
[1] IIIT Delhi, India; [2] NTU, Singpore; [3] IIT Delhi, India
{shivaniku, shad.akhtar}@iiitd.ac.in,cambria@ntu.edu.sg
tanchak@iitd.ac.in

Abstract

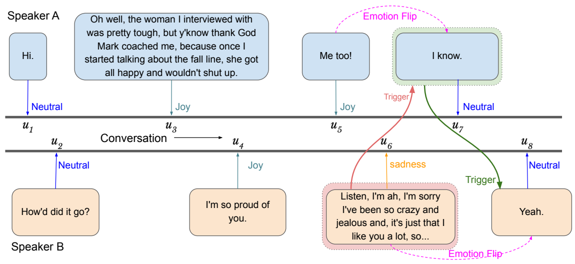
We present SemEval-2024 Task 10, a shared task centred on identifying emotions and finding the rationale behind their flips within monolingual English and Hindi-English code-mixed dialogues. This task comprises three distinct subtasks – emotion recognition in conversation for code-mixed dialogues, emotion flip reasoning for code-mixed dialogues, and emotion flip reasoning for English dialogues. Participating systems were tasked to automatically execute one or more of these subtasks. The datasets for these tasks comprise manually annotated conversations focusing on emotions and triggers for emotion shifts.[1] A total of 84 participants engaged in this task, with the most adept systems attaining F1-scores of 0.70, 0.79, and 0.76 for the respective subtasks. This paper summarises the results and findings from 24 teams alongside their system descriptions.

1 Introduction

In pursuit of one of AI's ultimate objectives, i.e., emulating human behaviour, machines must comprehend human emotions (Ekman, 1992; Picard, 1997). Consequently, Emotion Recognition in Conversation (ERC) has emerged as a vibrant domain within NLP (Hazarika et al., 2018b,a; Zhong et al., 2019; Ghosal et al., 2019; Jiao et al., 2020). The significance of emotion detection amplifies particularly during shifts in the speaker's emotional state. However, merely identifying an emotional transition is insufficient; understanding the catalyst behind the shift is crucial for facilitating informed decisions by other speakers. For instance, identifying the utterance responsible for a customer's transition from a positive emotional state (e.g., joy) to a negative one (e.g., disgust) due to a flawed dialogue system is critical in customer service. Such insights can serve as feedback to the dialogue system, enabling it to avoid (negative emotion-flip) or replicate (positive emotion-flip) similar utterances to enhance the customer experience in the future.



(a) Emotion-flip is caused by more than one utterance.



(b) Emotion-flip is caused by the previous utterance. Out of five emotion-flips, we show only two of them ($u_5 \rightarrow u_7$ and $u_6 \rightarrow u_8$) for brevity. Other emotion-flips are $u_1 \rightarrow u_3$, $u_2 \rightarrow u_4$, and $u_4 \rightarrow u_6$ with triggers u_3, u_3, and u_6, respectively.

Figure 1: Examples of emotion-flip reasoning.

Emotion-Flip Reasoning (EFR), as outlined by Kumar et al. (2021), presents a novel endeavour aimed at pinpointing the trigger utterances responsible for an emotion-flip within the context of a multi-party conversation. Figure 1 provides illustrative scenarios depicting the essence of EFR. In Figure 1a, Speaker B undergoes a transition in emotion (*neutral*→*fear*) between utterances u_6 and u_8. Notably, this emotional shift can be attributed to the contributions of Speaker A through utterances u_5 and u_7. The fundamental objective of this task is to discern these trigger utterances (u_5 and u_7) given a target emotion-flip utterance (u_8) and the context preceding it ($u_1 \ldots u_7$).

As an effort to advance research within the do-

[1]The task data is available at https://github.com/LCS2-IIITD/EDiReF-SemEval2024.git.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1933–1946
June 20-21, 2024 ©2024 Association for Computational Linguistics

main of ERC and EFR, this shared task at SemEval 2024 (Ojha et al., 2024) seeks to assess the efficacy of NLP systems in automatically addressing both of these tasks. Furthermore, as technological applications extend beyond English to encompass non-English, multilingual, and code-mixed populations, there is a growing need to broaden the scope of research. To support this cause and further the exploration of code-mixed languages, we advocate for the inclusion of ERC and EFR tasks within Hindi-English (Hinglish) code-mixed conversations. Specifically, the shared task is segmented into three distinct subtasks:

- ***Task A* – ERC in Hindi-English code-mixed conversation:** Given a multiparty code-mixed conversation, tag each utterance with one of the eight emotion labels – *anger*, *disgust*, *fear*, *sadness*, *surprise*, *joy*, *contempt*, and *neutral*.
- ***Task B* – EFR in Hindi-English code-mixed conversation:** Given a multiparty code-mixed conversation along with emotions for each utterance, the goal is to identify the trigger utterance for each emotion-flip in the dialogue.
- ***Task C* – EFR in English conversation:** It is similar to Task B but in monolingual English.

The decision to omit ERC for monolingual English stems from its thorough examination and the abundance of available datasets. Conversely, ERC in Hindi-English code-mixed conversation remains relatively unexplored, and as far as we are aware, no other dataset besides the one outlined in this article is publicly accessible.

Further elaboration on our task data and setting is provided in Sections 3 and Section 4, respectively. The participating teams are outlined in Section 5, with their task outcomes and assessments detailed in Section 6.

2 Related Work

Emotion recognition. Identifying emotions has been a focal point in prior research, with investigations into emotion analysis (Ekman, 1992; Picard, 1997; Cowen and Keltner, 2017; Mencattini et al., 2014; Zhang et al., 2016; Cui et al., 2020) initially centring on standalone inputs devoid of contextual cues. However, recognising the significance of contextual information, the emphasis shifted towards emotion detection within conversations, particularly ERC. Initially, ERC was tackled using heuristic approaches and conventional machine learning techniques (Fitrianie et al., 2003; Chuang and Wu, 2004; Li et al., 2007). However, the recent trend has witnessed a transition towards the adoption of a diverse array of deep learning methodologies (Hazarika et al., 2018a; Zhong et al., 2019; Li et al., 2020; Ghosal et al., 2019; Jiao et al., 2020; Hazarika et al., 2021; Shen et al., 2020; Poria et al., 2017; Jiao et al., 2019; Tu et al., 2022; Yang et al., 2022; Ma et al., 2022).

Emotion and code-mixing. Current studies addressing emotion analysis in code-mixed language primarily centre around isolated social media texts (Sasidhar et al., 2020; Ilyas et al., 2023; Wadhawan and Aggarwal, 2021) and reviews (Suciati and Budi, 2020; Zhu et al., 2022). Despite examinations into aspects like sarcasm (Kumar et al., 2022a,b), humour (Bedi et al., 2023), and offence (Madhu et al., 2023) within code-mixed conversations, the domain of emotion analysis remains largely uncharted, lacking pertinent literature, to the best of our knowledge. Our objective is to address this gap by delving into the under-explored realm of ERC, specifically within Hindi-English code-mixed dialogues in this shared task.

Beyond emotion recognition. The interpretability of emotion recognition within the linguistic domain represents a relatively uncharted avenue of research, with only a limited number of studies delving into this field. Previous works by Lee et al. (2010); Poria et al. (2021); Wang et al. (2023) have focused on investigating the root causes of expressed emotions in text, commonly referred to as 'emotion-cause analysis.' This task involves identifying a specific span within the text that elicits a particular emotion. While on an abstract level, both emotion-cause analysis and emotion-flip reasoning tasks may appear interconnected, they diverge significantly in practice. Emotion-cause analysis aims to pinpoint phrases within the text that provide clues or triggers for the expressed emotion. In contrast, our proposed EFR task pertains to conversational dialogues involving multiple speakers, with the objective of extracting the causes (Kumar et al., 2023a) or triggers behind emotional transitions for a speaker. The triggers comprise one or more utterances from the dialogue history, as illustrated in the two examples in Figure 1.

In this shared task, we tackle the challenge of automatically performing the task of ERC and EFR for code-mixed and monolingual English dialogues in order to further this research direction.

Split	Emotions							Total
	Disgust	Joy	Surprise	Anger	Fear	Neutral	Sadness	
Train	225	1466	1021	911	229	3702	576	8130
Dev	20	156	144	126	39	395	97	977
Test	61	325	238	283	42	943	169	2061

(a) ERC – English

Split	#D with Flip	#U with Flip	#Triggers
Train	834	4001	6740
Dev	95	427	495
Test	232	1002	1152

(b) EFR – English

Split	Emotions								Total
	Disgust	Joy	Surprise	Anger	Fear	Neutral	Sadness	Contempt	
Train	127	1646	444	856	530	4091	572	549	8815
Dev	21	242	68	122	91	652	132	75	1403
Test	21	382	57	150	129	697	167	87	1690

(c) ERC – Hindi

Split	#D with Flip	#U with Flip	#Triggers
Train	344	4406	5565
Dev	47	686	959
Test	58	781	1026

(d) EFR – Hindi

Table 1: Statistics of the English and Hindi datasets for ERC and EFR.

3 Data

English Conversations: We extend MELD (Poria et al., 2019), an established ERC dataset comprising monolingual English dialogues, by incorporating annotations for emotion-flip reasoning. These dialogues are sourced from the popular TV series *F.R.I.E.N.D.S*[2]. Each utterance u is attributed to a specific speaker s and assigned an emotion label e ∈[*anger*, *disgust*, *fear*, *sadness*, *surprise*, *joy*, *neutral*]. In the context of a speaker's emotional transition, we designate and label trigger utterances as 1 if they induce the speaker's emotional shift – the emotion alters from the speaker's preceding utterance within the same dialogue. In contrast, a label 0 indicates that the utterance bears no responsibility for the emotional transition.

To facilitate the annotation of triggers, we establish a set of guidelines outlined below. Within this framework, a ***trigger*** is defined as any utterance within the contextual history of the target utterance (the utterance for which the trigger is to be identified) meeting the following criteria:

1. An utterance, or part thereof, directly influencing a change in emotion of the target speaker is designated as the trigger.
2. The speaker of the trigger utterance may be different from or the same as the target speaker.
3. The target utterance itself may qualify as a trigger utterance if it contributes to the emotional transition of the target speaker. For instance, if an individual's emotion shifts from *neutral* to *sad* due to conveying a sad message, then the target utterance is deemed responsible for the transition.
4. Multiple triggers may be accountable for a single emotional transition.
5. In cases where the rationale behind an emotional transition is not identifiable from the data, no utterance should be labelled as a trigger.

In total, we have annotated emotion-flip reasoning for $1,161$ monolingual English conversation dialogues, encompassing $8,387$ trigger utterances across $5,430$ emotion-flip instances. Three annotators carefully annotated these dialogues in accordance with the aforementioned guidelines for trigger identification. Among the three, two annotators were male while one was female, all possessing 3-10 years of research experience within the 30-40 age bracket. We calculated the alpha-reliability inter-annotator agreement (Krippendorff, 2011) between each pair of annotators, yielding $\alpha_{AB} = 0.824$, $\alpha_{AC} = 0.804$, and $\alpha_{BC} = 0.820$. By averaging these scores, we derived an overall agreement score of $\alpha = 0.816$. We call the resultant dataset as `MELD-FR`.

Hindi-English Code-mixed Conversations: For code-mixed tasks, we adhere to identical guidelines as those applied to English, selecting code-mixed conversations from a preexisting dialogue dataset called MaSaC (Bedi et al., 2023). The dialogues in the dataset are sourced from the popular Indian TV series '*Sarabhai vs Sarabhai*'[3]. Further, we annotated $11,908$ utterances spanning 449 dialogues, encompassing eight emotion labels (including '*contempt*' alongside the six basic emotions and *neutral*) for the ERC task, achieving a Krippendorff alpha-reliability inter-annotator agreement (Krippendorff, 2011) of 0.85. In the context of EFR, we annotated $7,550$ trigger utterances for $5,873$ emotion-flip occurrences. Mirroring our approach with the English dataset, we engaged experts fluent in both Hindi and English to ensure accuracy. As

[2] https://www.imdb.com/title/tt0108778/

[3] https://www.imdb.com/title/tt1518542/

a measure of quality assurance, the Krippendorff alpha-reliability inter-annotator agreement stands at $\alpha = 0.853$. The resultant dataset is denoted as E-MASAC and EFR-MASAC for the ERC and EFR tasks, respectively. A concise overview of both datasets is presented in Table 1.

4 Task and Background

The idea of the presented shared task is to delve into ERC and EFR within the domain of English and code-mixed dialogues. This section delves into our preliminary investigations for the three subtasks entailed in this collaborative endeavour.

4.1 Shared Task Settings

Task A. In the task of ERC within code-mixed dialogues, participants receive textual utterances as input along with their respective speakers for each dialogue. Their objective is to develop systems capable of autonomously predicting the emotion labels for each utterance. Essentially, the system is presented with a dialogue $D_{erc} = \{(s_1, u_1), (s_2, u_2), ..., (s_n, u_n)\}$, and it must anticipate the emotions e_i for each utterance u_i uttered by speaker s_i. Weighted F-1 score of the emotion classification is used as the evaluation metric for the task of ERC.

Task B. For the code-mixed EFR task, participants receive dialogues along with their corresponding utterances, speakers, and emotions, presented in the format $D_{efr} = \{(s_1, u_1, e_1), (s_2, u_2, e_2), ..., (s_n, u_n, e_n)\}$. Their objective is to anticipate trigger utterances, T, from the context whenever a speaker undergoes an emotion flip. In other words, $T \in \{u_i, ..., u_j\}$ if $s_i = s_j$ and $e_i \neq e_j$. The evaluation metric of choice for this task is the F1 score obtained for trigger utterances.

Task C. The input modelling for Task C mirrors that of Task B, as both tasks revolve around EFR. Here, the data comes from the `MELD-FR` dataset and is present in Monolingual English. Just like Task B, the evaluation is conducted based on the F1 score achieved for trigger utterances.

4.2 Pilot Study

Task A. Our preliminary investigation for the task of ERC in code-mixed setting (Kumar et al., 2023b), we integrate commonsense knowledge with the dialogue representation acquired from a backbone architecture designed for dialogue understanding. We leverage the COMET graph (Bosselut et al., 2019) to extract commonsense knowledge, and subsequently employ context-aware attention (Yang et al., 2019) to integrate this information with the dialogue context. This adaptable module, when combined with RoBERTa (Liu et al., 2019), yields a weighted average F1-score of 0.44 in performance.

Task B. For evaluating the feasibility of our second subtask, we employ FastText multilingual word embeddings[4] for the tokens and perform classification using the proposed model for Task C to obtain an F1-score of 0.27 for trigger identification.

Task C. In our initial exploration (Kumar et al., 2021), we explored a memory-network and transformer-based architecture to address each occurrence of emotion-flip. This approach yielded a trigger-F1 score of 0.53. While these findings surpassed various baselines, the overall performance remains inadequate from a practical standpoint, with an error rate of approximately 50%.

5 Participants

A total of 84 participants engaged in the CodaLab competition organised for the shared task[5], with 24 teams (Shaik et al., 2024; V et al., 2024; Liang et al., 2024; Venkatesh et al., 2024; Yenumulapalli et al., 2024; Niță and Păiș, 2024; Moctezuma et al., 2024; Vyas, 2024; Tareh et al., 2024; Garcia et al., 2024; Wan et al., 2024; Abootorabi et al., 2024; Patel et al., 2024; Siino, 2024; Nguyen and Zhang, 2024; Vaidya et al., 2024; Takahashi, 2024; Alexandru et al., 2024; Rajesh et al., 2024; Shanbhag et al., 2024; Creanga and Dinu, 2024; pan et al., 2024) submitting papers describing their systems. Among the submissions, a prevailing trend emerges with the widespread adoption of Large Language Models (LLMs) such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT (Radford et al., 2019), LLaMa (Touvron et al., 2023), and Mistral (Jiang et al., 2023). Techniques such as fine-tuning, instruction tuning, ensembling, and prompting significantly contribute to enhanced performance in the task. Moreover, many approaches utilise machine learning-based methods including linear regression and SVM. Additionally, some studies explore statistical and rule-based methods such as TF-IDF. While LLMs dominate the approaches for both ERC and EFR, machine learning methods

[4] `https://fasttext.cc/docs/en/crawl-vectors.html`

[5] `https://codalab.lisn.upsaclay.fr/competitions/16769`

also remain popular among the participants. Furthermore, there appears to be a notable preference among teams for Task A over Tasks B and C, as evidenced by the higher participation in Task A compared to the latter two. An overview of the top-performing models from various teams for ERC is provided in Table 2, while Table 3 presents the systems for EFR. We summarize some of the techniques used by the top performing systems below.

Using LLMs There exists a prevalent preference for LLMs among teams addressing the ERC and EFR tasks, with approximately 18 methods leveraging LLMs for these endeavours. Notably, BERT and its variants emerge as the most favoured models. Some teams explore larger open-source language models like Zephyr (Tunstall et al., 2023) and Mistral, while at least one team delves into closed-source alternatives such as GPT3.5 (Brown et al., 2020). In the realm of ERC, the leading system (refer to Table 4) integrates DistilBERT (Sanh et al., 2020) with classical machine learning techniques to execute emotion classification optimally. Although the authors experiment with BERT, RoBERTa, and GPT-4 (OpenAI et al., 2023), their most effective model combines DistilBERT with classical ML algorithms. They adopt a two-step approach, initially extracting contextual features from dialogues using an LLM, then inputting these features into classical ML algorithms such as random forests, SVM, logistic regression, and Naive Bayes. Notably, DistilBERT outperforms GPT-4, possibly attributed to the latter's extensive parameter count, necessitating substantial data for meaningful learning. However, our Task A dataset (E-MASAC) encompasses only $\sim$ 8500 utterances, limiting the efficacy of larger models. Conversely, lighter models like DistilBERT exhibit superior adaptability with limited data, capturing nuanced patterns effectively. This finding aligns with observations from various teams, including BITS Pilani, where BERT outperforms Llama.

For the task of EFR as well, LLMs appear to be the predominant choice among the teams. However, intriguingly, the most effective model for this task (refer to Table 5) adopts a classical machine learning approach - XGBoost. Further elaboration on this aspect is provided in Section 6.2.

Classical machine learning and deep learning methods Efficiently capturing context information is crucial in modelling conversations. Several teams explored this aspect, utilising Recurrent

Team Name	Backbone Architecture	Model Type
AIMA	GPT3.5 + ML	Ensemble
BITS Pilani	BERT	LLM
CLTeam1	RoBERTa & BERT	LLM+Ensemble
FeedForward	Zephyr	LLM
Hidetsune	SpaCy-v3	ML
IASBS	DistilBERT + ML	LLM+ML
IITK	Transformer + GRU	LLM+DL
INGEOTEC	Bag of Words	Statistical
Innovators	SVM	ML
ISDS-NLP	RoBERTa	LLM
MorphingMinds	LR	ML
RACAI	BERT + ML	LLM+ML
SSN_ARMM	TF-IDF	Statistical
SSN_Semeval10	BERT	LLM
TECHSSN	LSTM	DL
TECHSSN1	RoBERTa	LLM
TransMistral	Mistral 7B	LLM
TW-NLP	MBERT	LLM
UCSC NLP	BERT	LLM
UMUTeam	BERT	LLM
VerbaNexAI Lab	Transformer + GRU	LLM+DL
YNU-HPCC	DeBERTa	LLM

Table 2: Summary of the models according to the submitted system descriptions for Task A (ERC).

Team Name	Backbone Architecture	Model Type
FeedForward	Zephyr	LLM
GAVx	GPT3.5	LLM
IASBS	DistilBERT + ML	LLM+ML
IITK	Transformer + GRU	LLM+DL
Innovators	-	Rule Based
LinguisTech	-	NER Model
SSN_ARMM	TF-IDF	Statistical
TECHSSN	LSTM	DL
TW-NLP	XGBoost	ML
UCSC NLP	BERT + GRU	LLM+DL
UMUTeam	BERT	LLM
YNU-HPCC	DeBERTa	LLM

Table 3: Summary of the models according to the submitted system descriptions for Task B and C (EFR).

Neural Networks (RNNs) like LSTMs and GRUs. Specifically, at least three teams have integrated GRU with Transformers to enhance context capture. Conversely, team TECHSSN adopts a simpler approach, employing LSTM with intelligent embedding layers for both ERC and EFR tasks. However, these methods frequently fall short in comparison to utilising pre-trained LLMs, as outlined in Section 6.2.

Rule-based and statistical methods The surge in deep learning's popularity can be attributed to the remarkable advancements in LLMs. This has led to a decline in the usage of traditional rule-based or statistical approaches, despite their potential to perform comparably in certain scenarios

Rank	Team Name	Results
3	IASBS (Tareh et al., 2024)	0.70
5	FeedForward (Shaik et al., 2024)	0.51
6	TW-NLP (Tian et al., 2024)	0.46
7	TECHSSN1 (Yenumulapalli et al., 2024)	0.45
9	IITK (Patel et al., 2024)	0.45
10	UCSC NLP (Wan et al., 2024)	0.45
11	CLTeam1 (Vaidya et al., 2024)	0.44
13	UMUTeam (pan et al., 2024)	0.43
12	ISDS-NLP (Creanga and Dinu, 2024)	0.43
14	BITS Pilani (Venkatesh et al., 2024)	0.42
15	AIMA (Abootorabi et al., 2024)	0.42
16	SSN_Semeval10 (Rajesh et al., 2024)	0.40
17	Hidetsune (Takahashi, 2024)	0.39
18	INGEOTEC (Moctezuma et al., 2024)	0.39
19	SSN_ARMM (S et al., 2024)	0.38
20	TransMistral (Siino, 2024)	0.36
23	TECHSSN (V et al., 2024)	0.34
24	MorphingMinds (Vyas, 2024)	0.33
26	RACAI (Niță and Păiș, 2024)	0.31
27	Innovators (Shanbhag et al., 2024)	0.28
30	VerbaNexAI Lab (Garcia et al., 2024)	0.24
32	YNU-HPCC (Liang et al., 2024)	0.18

Table 4: Results (Weighted F1) for Task A. Rank is as mentioned in CodaLab. Team Name is as mentioned in the corresponding system description.

alongside more intricate machine learning or deep learning methods. It was pleasant to observe numerous teams incorporating such traditional techniques into this shared task. Notably, at least four teams opted for methods like Bag of Words, TF-IDF, NER based, and rule based approaches. While these methods may not excel in the ERC task, they surprisingly demonstrate superiority in the EFR task. This is reasoned in detail in Section 6.2.

6 Results

In this section, we delve into the outcomes achieved by the participating teams in the shared task outlined earlier. Initially, we will examine the results submitted by the 24 teams, which provided detailed descriptions of their systems. Subsequently, we will present the leaderboard, showcasing the performance rankings of all participants.

6.1 Task A: ERC in Hindi-English code-mixed conversation

The results for Task A are compiled in Table 4. Out of the 24 submitted papers, 22 teams explored the code-mixed ERC task, attaining weighted F1 scores spanning from 0.70 to 0.18. Notably, the foremost twelve teams, up to SSN_Semeval10, opted for LLMs as their architectural preference, yielding top performances. Following closely, RNN-based approaches such as LSTM and classical ML methods like SVM emerged as the subsequent choices. A notable observation is the substantial disparity (approximately 37%) in performance between the leading model and the succeeding system.

Both leading teams relied on LLMs as their primary architectural framework, yet IASBS diverged by integrating classical ML methods. Their innovative two-phase strategy, combining LLMs for contextual representations and ML techniques for classification, evidently yielded significant improvements. Conversely, the subsequent top model utilised LLMs without any ensembling. Team FeedForward, securing fifth place on the CodaLab leaderboard for Task A, implemented instruction-based finetuning and quantized low-rank adaptation alongside novel techniques like sentext-height and enhanced prompting strategies.

Another intriguing observation arises from the marginal discrepancy (approximately 2%) between strategies based on LLMs and those employing classical ML techniques. Team SSN_Semeval10 refined a BERT classifier, achieving a weighted F1-score of 0.40. Conversely, team Hidetsune took a different approach by translating all code-mixed data into English and employing data augmentation to bolster the 'English'-based ERC dataset. Subsequently, they trained a SpaCy-v3[6] classifier, resulting in a weighted F1-score of 0.39.

Rank	Team Name	Results
1	TW-NLP (Tian et al., 2024)	0.79
2	Innovators (Shanbhag et al., 2024)	0.79
2	UCSC NLP (Wan et al., 2024)	0.79
2	GAVx (Nguyen and Zhang, 2024)	0.79
3	FeedForward (Shaik et al., 2024)	0.77
5	IITK (Patel et al., 2024)	0.56
6	UMUTeam (pan et al., 2024)	0.26
7	IASBS (Tareh et al., 2024)	0.12
9	SSN_ARMM (S et al., 2024)	0.11
11	TECHSSN (V et al., 2024)	0.1
21	YNU-HPCC (Liang et al., 2024)	0.01

Table 5: Results for Task B. F1 score for trigger utterances is our metric of choice. Rank is as mentioned in CodaLab. Team Name is as mentioned in the corresponding system description.

6.2 Task B – EFR in Hindi-English code-mixed conversation

Table 5 presents the outcomes for Task B, wherein the highest performance attained a trigger F1-score

[6]https://spacy.io/

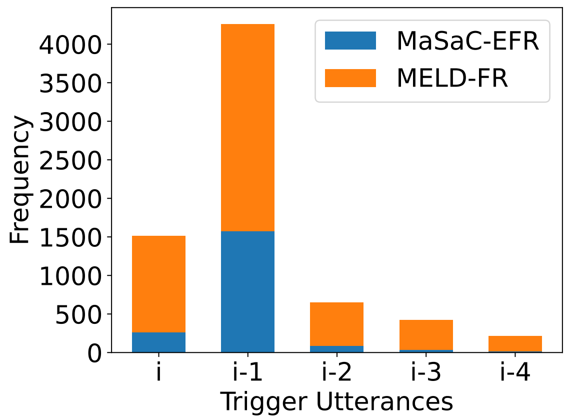


Figure 2: Distribution of triggers for the last four utterances from the trigger utterance i.

of 0.79. Particularly intriguing is the fact that the leading four teams achieved identical F1-scores, with the top two teams opting for conventional ML and rule-based approaches. This phenomenon stems from the common occurrence wherein a speaker's emotional shift in a conversation at utterance i is predominantly triggered by the $i-1$ utterance. This pattern underscores the significance of the preceding utterance as a trigger. Illustrated in Figure 2 is the trigger distribution within the dialogues of EFR-MASAC and MELD-FR. Evidently, the majority of trigger utterances are the $i-1^{th}$ utterances. Employing XGBoost for trigger classification, the leading team, TW-NLP, secured their position, while the second-ranking team opted for a rule-based approach, designating all $i-1$ utterances as triggers. This strategy led to the attainment of the highest score of 0.79 F1.

6.3 Task C – EFR in English conversation

The outcomes for Task C are displayed in Table 6, revealing the top-performing system achieving an F1 score of 0.76 for the triggers. Impressively, the subsequent results closely trail the best one, exhibiting only a marginal gap of approximately $2\% to 4\%$. Notably, the leading two performers in the task predominantly utilise methods employing LLMs, while the third-best performance is attributed to XGBoost. Illustrated in Figure 2, MELD-FR also grapples with a skewed distribution of trigger utterances, thereby resulting in comparable performances between LLMs and ML-based systems.

6.4 Findings by Participants

Challenge of code-mixing. The dataset utilised in this shared task encompasses Hindi-English code-mixed instances for subtasks A and B, presenting the most formidable challenge of the competition. To address this hurdle, several teams, including TransMistral, FeedForward, and Hidetsune, opted for translation, converting all code-mixed instances into monolingual English before engaging in any classification process. Additionally, teams such as TW-NLP leveraged multilingual LLMs like MBERT to effectively manage code-mixed input.

Effect of data augmentation. Machine learning and deep learning techniques exhibit an insatiable appetite for data, giving rise to circumstances where an abundance of data tends to correlate with improved performance. In light of this conjecture, several teams, including Hidetsune, ventured into experimenting with data augmentation for the ERC task. The general observation revealed an enhancement in performance with the incorporation of more data during model training.

Required context for classification. Emotions are fleeting and are typically influenced by the immediate circumstances in which the speaker finds themselves. As a result, the nearby utterances within a dialogue exert a more pronounced impact on determining the emotional nuances of a speaker compared to utterances further removed in context. This phenomenon is depicted in Figure 2. Consequently, teams such as FeedForward and IITK initially ascertain the requisite extent of context needed for conducting ERC, before proceeding with classification, taking the computed context into consideration.

Challenge of implicit triggers. Emotion flips can generally be attributed to two scenarios: firstly, when something uttered in the dialogue directly prompts the emotion flip, constituting explicit triggers; and secondly, when events external to the

Rank	Team Name	Results
2	GAVx (Nguyen and Zhang, 2024)	0.76
3	FeedForward (Shaik et al., 2024)	0.74
5	TW-NLP (Tian et al., 2024)	0.71
7	Innovators (Shanbhag et al., 2024)	0.68
8	UCSC NLP (Wan et al., 2024)	0.68
10	IITK (Patel et al., 2024)	0.6
11	SSN_ARMM (S et al., 2024)	0.26
12	IASBS (Tareh et al., 2024)	0.25
13	TECHSSN (V et al., 2024)	0.24
15	UMUTeam (pan et al., 2024)	0.22
26	YNU-HPCC (Liang et al., 2024)	0.07

Table 6: Results for Task C. F1 score for trigger utterances is our metric of choice. Rank is as mentioned in CodaLab. Team Name is as mentioned in the corresponding system description.

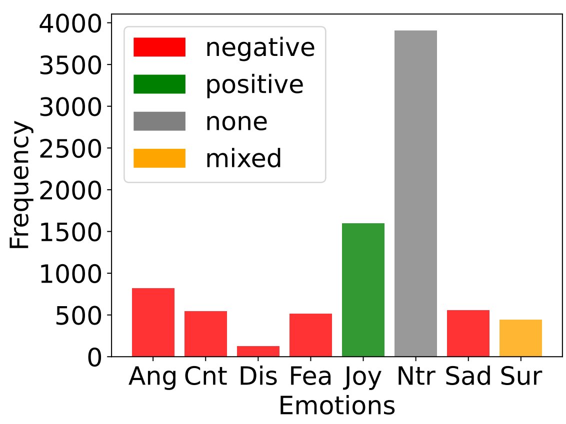


Figure 3: Emotion distribution in E-MASAC. The colors depict the distribution of emotions capturing positive, negative, mixed, and no feelings (Abbreviations: Ang: Anger, Cnt: Contempt, Dis: Disgust, Fea: Fear, Ntr: Neutral, Sad: Sadness, Sur: Surprise).

dialogue, such as an act of theft, occur without explicit mention in the dialogue, representing implicit triggers. In both the EFR-MASAC and MELD-FR datasets, instances of implicit triggers exist where no trigger utterances are marked in the dialogue. These instances present a challenge for the learned models of several teams, including GAVx.

Negative vs positive emotions. The dataset E-MASAC utilises Ekman emotions (Ekman, 1992) as its set of emotion labels, encompassing six emotions and one label for neutral emotions. These emotions include Anger, Contempt, Disgust, Fear, Joy, Neutral, Sadness, and Surprise. Notably, among these emotions, five portray negative feelings (Anger, Contempt, Disgust, Fear, and Sadness), while only one represents positive emotions (Joy). Surprise, on the other hand, can convey either positive or negative emotions depending on the context. Figure 3 displays the distribution of these emotions within E-MASAC. It's evident that as there's only one category for positive emotions, all such instances are classified as joy, leading to a higher frequency of joy compared to other emotions. Moreover, the neutral category has the most instances compared to the others. Consequently, many teams, like IITK, have noted that their models perform better for the neutral and joy labels than for any other emotion.

6.5 Leaderboard

In this paper, we have exclusively examined the outcomes of participants who provided a description of their system(s) for the shared task. The complete array of ranks, team names, and results

Team	Task A	Task B	Task C
MasonTigers	0.78 (1)	0.79 (2)	0.79 (1)
Knowdee	0.73 (2)	0.66 (4)	0.61 (9)
IASBS	0.70 (3)	0.12 (7)	0.25 (12)
-	0.66 (4)	0.07 (20)	0.04 (28)
FeedForward	0.51 (5)	0.77 (3)	0.74 (3)
TW-NLP	0.45 (6)	0.79 (1)	0.71 (5)
TechSSN1	0.45 (7)	0.00 (22)	0.00 (29)
-	0.45 (8)	0.79 (2)	0.68 (8)
IITK	0.45 (9)	0.56 (5)	0.60 (10)
-	0.45 (10)	0.10 (11)	0.15 (23)
CLTeam1	0.44 (11)	0.10 (11)	0.24 (13)
UniBucNLP	0.43 (12)	0.10 (14)	0.06 (27)
UMUTeam	0.43 (13)	0.26 (6)	0.22 (15)
BITS Pilani	0.42 (14)	0.10 (11)	0.24 (13)
AIMA	0.42 (15)	0.10 (12)	0.21 (21)
SSN_Semeval10	0.40 (16)	0.00 (22)	0.00 (29)
OZemi	0.39 (17)	0.00 (22)	0.00 (29)
INGEOTEC	0.39 (18)	0.00 (22)	0.00 (29)
-	0.38 (19)	0.11 (9)	0.26 (11)
-	0.37 (20)	0.07 (19)	0.07 (25)
CUET_NLP	0.37 (21)	0.00 (22)	0.00 (29)
-	0.36 (22)	0.10 (10)	0.22 (19)
TechSSN	0.34 (23)	0.10 (11)	0.24 (13)
MorphingMinds	0.33 (24)	0.10 (16)	0.22 (18)
Z-AGI Labs	0.31 (25)	0.00 (22)	0.00 (29)
RACAI	0.31 (26)	0.10 (11)	0.24 (13)
Innovators	0.28 (27)	0.79 (2)	0.68 (7)
-	0.27 (28)	0.10 (13)	0.21 (20)
-	0.26 (29)	0.10 (15)	0.16 (22)
-	0.24 (30)	0.00 (22)	0.74 (4)
LinguisTech	0.24 (30)	0.00 (22)	0.70 (6)
Team + 1	0.24 (30)	0.09 (17)	0.22 (16)
PartOfGlitch	0.24 (30)	0.00 (22)	0.10 (24)
VerbNexAI Lab	0.24 (30)	0.00 (22)	0.00 (29)
-	0.24 (30)	0.00 (22)	0.00 (29)
silp_nlp	0.24 (31)	0.11 (8)	0.23 (14)
-	0.18 (32)	0.01 (21)	0.07 (26)
-	0.14 (33)	0.09 (18)	0.22 (17)
GAVx	0.08 (34)	0.79 (2)	0.76 (2)

Table 7: Leaderboard from CodaLab. Rank for each task is mentioned in parenthesis. Top three systems are highlighted in green, yellow, and orange.

of the participants in the CodaLab competition, encompassing users who didn't submit a system description, is depicted in Table 7. Although not all teams attempted all three tasks, however, they obtained some score in the leaderboard since they submitted some values for the output. A glimpse at the leaderboard reveals that the best performance for tasks A, B, and C stood at 0.78, 0.79, and 0.79, respectively. One, four, and one team(s) achieved best performance of the three tasks.

7 Conclusion

This paper outlines SemEval 2024 Task 10, covering its goals, data, participants, and results. It includes three subtasks: emotion identification in code-mixed dialogues and pinpointing triggers for emotion shifts in code-mixed and English dialogues. 84 participants competed on CodaLab, with 24 teams submitting system description papers. Top systems for Tasks A and C used LLM-based architectures, while Task B favored standard ML techniques. Leading systems achieved F1 scores of 0.70, 0.79, and 0.76 across subtasks, indicating impressive performance but also highlighting ongoing challenges for future research.

References

Mohammad Mahdi Abootorabi, Nona Ghazizadeh, Seyed Arshan Dalili, Alireza Ghahramani Kure, Mahshid Dehghani, and Ehsaneddin Asgari. 2024. Aima at semeval-2024 task 10: History-based emotion recognition in hindi-english code-mixed conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1715–1721, Mexico City, Mexico. Association for Computational Linguistics.

Mihaela Alexandru, Călina Georgiana Ciocoiu, Ioana Măniga, Octavian Ungureanu, Daniela Gîfu, and Diana Trandăbăț. 2024. Linguistech at semeval-2024 task 10: Emotion discovery and reasoning its flip in conversation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 399–406, Mexico City, Mexico. Association for Computational Linguistics.

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Trans. Affect. Comput.*, 14(2):1363–1375.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

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

Ze-Jing Chuang and Chung-Hsien Wu. 2004. Multi-modal emotion recognition from speech and text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, pages 45–62.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *PNAS*, 114(38):E7900–E7909.

Claudiu Creanga and Liviu P. Dinu. 2024. Isds-nlp at semeval-2024 task 10: Transformer based neural networks for emotion recognition in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 636–641, Mexico City, Mexico. Association for Computational Linguistics.

Heng Cui, Aiping Liu, Xu Zhang, Xiang Chen, Kongqiao Wang, and Xun Chen. 2020. Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge-Based Systems*, 205:106243.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Siska Fitrianie, Pascal Wiggers, and Leon JM Rothkrantz. 2003. A multi-modal eliza using natural language processing and emotion recognition. In *Text, Speech and Dialogue: 6th International Conference, TSD 2003, České Budějovice, Czech Republic, September 8-12, 2003. Proceedings 6*, pages 394–399. Springer.

Santiago Garcia, Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez-Santos, and Edwin Puertas. 2024. Verbanexai lab at semeval-2024 task 10: Emotion recognition and reasoning in mixed-coded conversations based on an nrc vad approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1322–1328, Mexico City, Mexico. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2021. Conversational transfer learning for emotion recognition. *Information Fusion*, 65:1–12.

Abdullah Ilyas, Khurram Shahzad, and Muhammad Kamran Malik. 2023. Emotion detection in code-mixed roman urdu - english text. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

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.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *AAAI*, volume 34, pages 8002–8009.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. *arXiv preprint arXiv:1904.04446*.

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

Shivani Kumar, Shubham Dudeja, Md Shad Akhtar, and Tanmoy Chakraborty. 2023a. Emotion flip reasoning in multiparty conversations. *IEEE Transactions on Artificial Intelligence*, pages 1–10.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022a. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.

Shivani Kumar, Ishani Mondal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022b. Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues.

Shivani Kumar, Ramaneswaran S, Md Akhtar, and Tanmoy Chakraborty. 2023b. 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. 2021. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *arXiv 2103.12360 (cs.CL)*.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA.

Haifang Li, Na Pang, Shangbo Guo, and Heping Wang. 2007. Research on textual emotion recognition incorporating personality factor. In *2007 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2222–2227. IEEE.

Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2020. Bieru: bidirectional emotional recurrent unit for conversational sentiment analysis. *arXiv preprint arXiv:2006.00492*.

Chenyi Liang, Jin Wang, and Xuejie Zhang. 2024. Ynu-hpcc at semeval-2024 task10: Pre-trained language model for emotion discovery and reasoning its flip in conversation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 764–771, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hui Ma, Jian Wang, Hongfei Lin, Xuejun Pan, Yijia Zhang, and Zhihao Yang. 2022. A multi-view network for real-time emotion recognition in conversations. *Knowledge-Based Systems*, 236:107751.

Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications*, 215:119342.

Arianna Mencattini, Eugenio Martinelli, Giovanni Costantini, Massimiliano Todisco, Barbara Basile, Marco Bozzali, and Corrado Di Natale. 2014. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems*, 63:68–81.

Daniela Moctezuma, Eric Tellez, Jose Ortiz Bejar, and Mireya Paredes. 2024. Ingeotec at semeval-2024 task 10: Bag of words classifiers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1104–1109, Mexico City, Mexico. Association for Computational Linguistics.

Vy Nguyen and Xiuzhen Zhang. 2024. Gavx at semeval-2024 task 10: Emotion flip reasoning via stacked instruction finetuning of llms. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 319–329, Mexico City, Mexico. Association for Computational Linguistics.

Sara Niță and Vasile Păiș. 2024. Racai at semeval-2024 task 10: Combining algorithms for code-mixed emotion recognition in conversation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1021–1026, Mexico City, Mexico. Association for Computational Linguistics.

Atul Kr. Ojha, Harish Tayyar Madabushi, Giovanni Da San Martino, A. Seza Doğruöz, Sara Rosenthal, and Aiala Rosá, editors. 2024. *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics, Mexico City, Mexico.

OpenAI, :, and Josh Achiam et al. 2023. Gpt-4 technical report.

ronghao pan, José Antonio García-Díaz, Diego Roldán, and Rafael Valencia-García. 2024. Umuteam at semeval-2024 task 10: Discovering and reasoning about emotions in conversation using transformers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 689–695, Mexico City, Mexico. Association for Computational Linguistics.

Shubham Patel, Divyaksh Shukla, and Ashutosh Modi. 2024. Iitk at semeval-2024 task 10: Who is the speaker? improving emotion recognition and flip reasoning in conversations via speaker embeddings. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1823–1832, Mexico City, Mexico. Association for Computational Linguistics.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*, pages 873–883.

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

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Antony Rajesh, Supriya Abirami A, Aravindan Chandrabose, and Senthil Kumar B. 2024. Ssn_semeval10 at semeval-2024 task 10: Emotion discovery and reasoning its flip in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 540–544, Mexico City, Mexico. Association for Computational Linguistics.

Rohith Arumugam S, Angel Deborah S, Rajalakshmi Sivanaiah, Milton R S, and Mirnalinee ThankaNadar. 2024. Ssn_armm at semeval-2024 task 10: Emotion recognition and trigger detection in code-mixed dialogues. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 716–722, Mexico City, Mexico. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

T Tulasi Sasidhar, Premjith B, and Soman K P. 2020. Emotion detection in hinglish(hindi+english) code-mixed social media text. *Procedia Computer Science*, 171:1346–1352. Third International Conference on Computing and Network Communications (CoCoNet'19).

Zuhair Hasan Shaik, Dhivya Prasanna R, Enduri Jahnavi, Rishi Thippireddy, VAMSI MADHAV P S S, SUNIL SAUMYA, and Shankar Biradar. 2024. Feedforward at semeval-2024 task 10: Trigger and sentext-height enriched emotion analysis in multi-party conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 731–742, Mexico City, Mexico. Association for Computational Linguistics.

Abhay Shanbhag, Suramya Jadhav, Shashank Rathi, Siddhesh Pande, and Dipali Kadam. 2024. Innovators at semeval-2024 task 10: Revolutionizing emotion recognition and flip analysis in code-mixed texts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 621–628, Mexico City, Mexico. Association for Computational Linguistics.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *arXiv preprint arXiv:2012.08695*.

Marco Siino. 2024. Transmistral at semeval-2024 task 10: Using mistral 7b for emotion discovery and reasoning its flip in conversation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 297–303, Mexico City, Mexico. Association for Computational Linguistics.

Andi Suciati and Indra Budi. 2020. Aspect-based sentiment analysis and emotion detection for code-mixed review. *International Journal of Advanced Computer Science and Applications*, 11(9).

Hidetsune Takahashi. 2024. Hidetsune at semeval-2024 task 10: An english based approach to emotion recognition in hindi-english code-mixed conversations using machine learning and machine translation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 367–371, Mexico City, Mexico. Association for Computational Linguistics.

Mehrzad Tareh, Aydin Mohandesi, and Ebrahim Ansari. 2024. Iasbs at semeval-2024 task 10: Delving into emotion discovery and reasoning in code-mixed conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1219–1228, Mexico City, Mexico. Association for Computational Linguistics.

Wei Tian, PeiYu Ji, Lei Zhang, and Yue Jian. 2024. Tw-nlp at semeval-2024 task10: Emotion recognition and emotion reversal inference in multi-party dialogues. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 304–308, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. 2022. Context- and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*, 3(5):699–708.

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.

Ravindran V, Shreejith G, Aashika Jetti, Rajalakshmi Sivanaiah, Angel Deborah S, Mirnalinee ThankaNadar, and Milton R S. 2024. Techssn at semeval-2024 task 10: Lstm-based approach for emotion detection in multilingual code-mixed conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 749–756, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Vaidya, Aditya Gokhale, Arnav Desai, Ishaan Shukla, and Sheetal Sonawane. 2024. Clteam1 at semeval-2024 task 10: Large language model based ensemble for emotion detection in hinglish. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 358–362, Mexico City, Mexico. Association for Computational Linguistics.

Dilip Venkatesh, Pasunti Prasanjith, and Yashvardhan Sharma. 2024. Bits pilani at semeval-2024 task 10: Fine-tuning bert and llama 2 for emotion recognition in conversation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 798–802, Mexico City, Mexico. Association for Computational Linguistics.

Monika Vyas. 2024. Morphingminds at semeval-2024 task 10: Emotion recognition in conversation in hindi-english code-mixed conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1207–1211, Mexico City, Mexico. Association for Computational Linguistics.

Anshul Wadhawan and Akshita Aggarwal. 2021. Towards emotion recognition in hindi-english code-mixed data: A transformer based approach.

Neng Wan, Steven Au, Esha Ubale, and Decker Krogh. 2024. Ucsc nlp at semeval-2024 task 10: Emotion discovery and reasoning its flip in conversation (ediref). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1503–1508, Mexico City, Mexico. Association for Computational Linguistics.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Baosong Yang, Jian Li, Derek F. Wong, Lidia S. Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):387–394.

Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of* *the AAAI Conference on Artificial Intelligence*, volume 36, pages 11595–11603.

Venkatasai Ojus Yenumulapalli, Pooja Premnath, Parthiban Mohankumar, Rajalakshmi Sivanaiah, and Angel Deborah S. 2024. Techssn1 at semeval-2024 task 10: Emotion classification in hindi-english code-mixed dialogue using transformer-based models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 820–825, Mexico City, Mexico. Association for Computational Linguistics.

Li Zhang, Kamlesh Mistry, Siew Chin Neoh, and Chee Peng Lim. 2016. Intelligent facial emotion recognition using moth-firefly optimization. *Knowledge-Based Systems*, 111:248–267.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China.

Xun Zhu, Yinxia Lou, Hongtao Deng, and Donghong Ji. 2022. Leveraging bilingual-view parallel translation for code-switched emotion detection with adversarial dual-channel encoder. *Knowledge-Based Systems*, 235:107436.