Emotion Discovery and Reasoning its Flip in Conversation

Final Presentation

Course: Statistical Natural Language Processing (CS779A)
Instructor: Ashutosh Modi

Group No: 08
Members: Sachin Bhadang (200831)
Yashwant Mahajan (201156)
Ritam Acharya (210859)

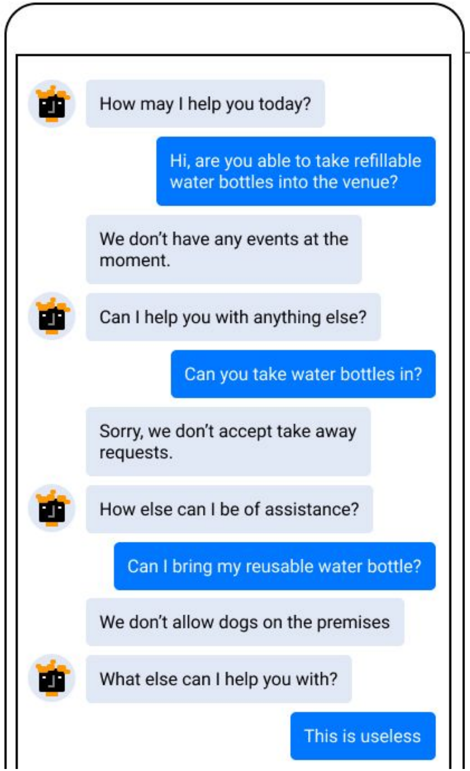
Motivation for Emotion Recognition (ERC)

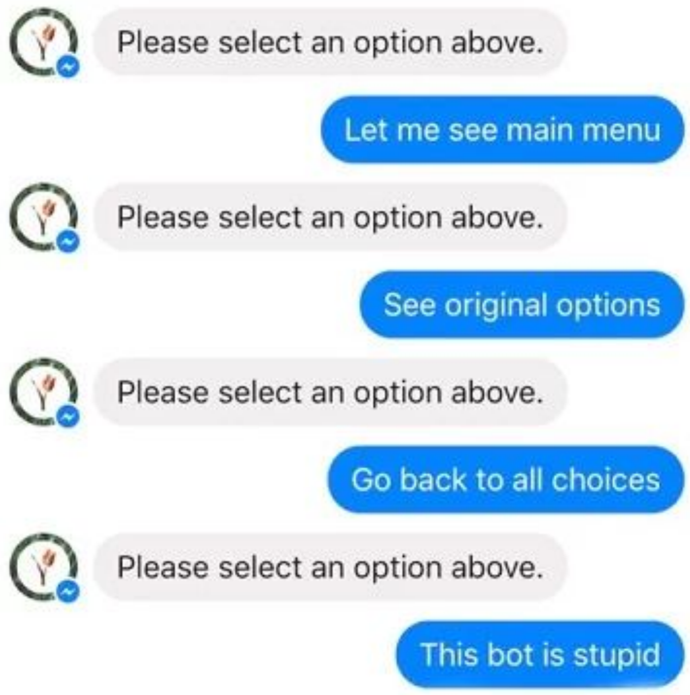
- Understanding Multilingual Emotions
- Enhancing User Experience
- Supporting Mental Health
- Navigating Shifting Sentiments

Motivation for Emotion Flip Reasoning (EFR)

- Identifying emotions in a conversation is not sufficient to understand the emotional dynamics of a speaker.
- Emotion-flip reasoning is a proposed task that aims to address this gap.
- It involves using flip explanations via triggers and instigators to examine how specific remarks or expressions influence the listener.
- These triggers and instigators can provide feedback for a response generation mechanism

Examples of Emotion Flip Reasoning (EFR)





Problem Statement

The problem contains 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

(iii) EFR in English conversations.

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

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

State-Of-The-Art for ERC

- **InstructERC** (*Lei, 2023*):
  - Reformulates the ERC task from a discriminative framework to a generative framework based on Large Language Models (LLMs)

- **SPCL-CL-ERC** (*Song, 2022*):
  - Supervised Prototypical Contrastive Learning (SPCL) loss
  - Integrates Prototypical Network and supervised contrastive learning

- **HiDialog** (*Liu, 2023*):
  - Hierarchical Dialogue Understanding model, insert multiple special tokens into a dialogue and propose the turn-level attention to learn turn embeddings hierarchically.
  - Then, a heterogeneous graph module is leveraged to polish the learned embeddings

State-Of-The-Art for ERC on MELD Dataset

Model	Weighted-F1
InstructERC (Lei, 2023)	69.15
SPCL-CL-ERC (Song, 2022)	67.25
HiDialog (Liu, 2023)	66.96

Model	Accuracy
M2FNet (V. Chudasama, 2022)	67.85
CFN-ESA (J Li, 2023)	67.85
SACL_LSTM (D Hu, 2023)	67.51

https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld

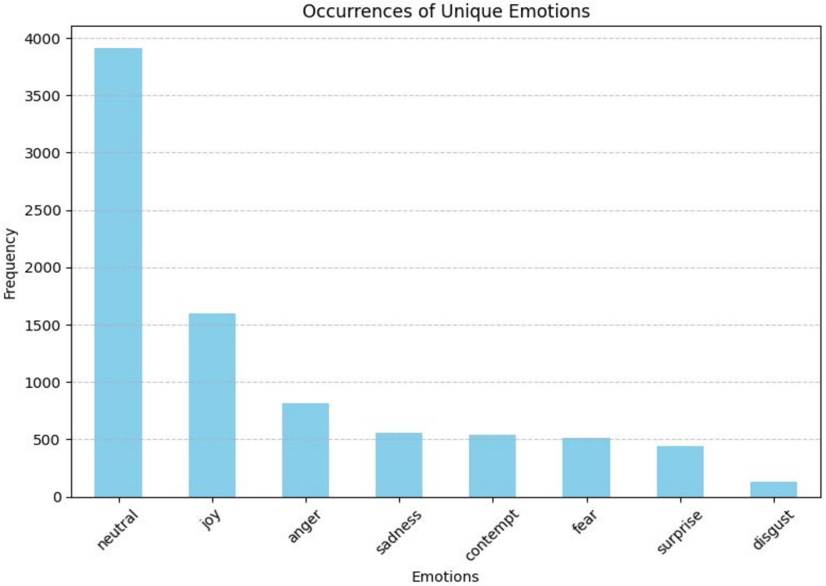
Data Analysis

- We have been provided with different datasets for each task
- The datasets are extracted from TV shows
  - MaSaC dataset from “Sarabhai vs Sarabhai”
  - MELD dataset from “Friends”
- Data contains -
  - Episode no.
  - Speaker name
  - Utterances
  - Emotions
  - Triggers (except task 1 data)

Analysis of Data for Task 1
(ERC in Hindi-English mixed)

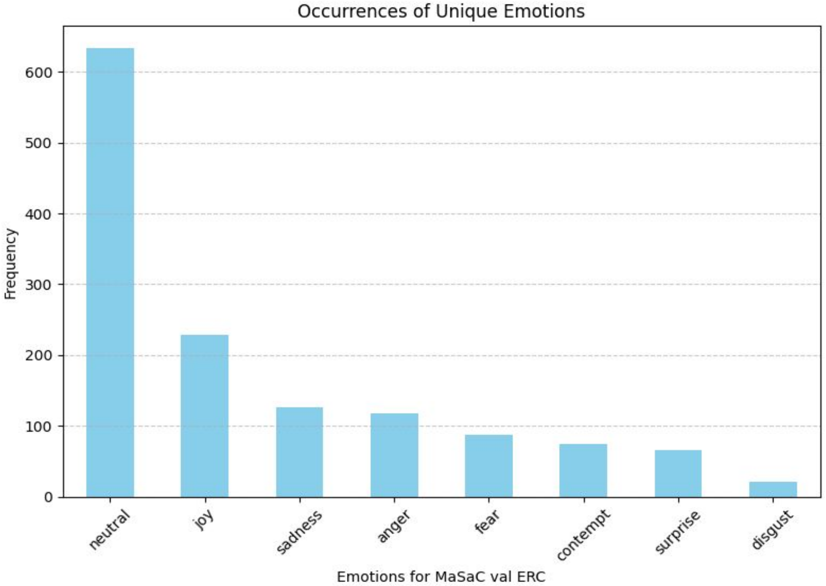
Train Data: No. of Episodes: 343

No. of Sentences: 8506



Val Data: No. of Episodes: 46

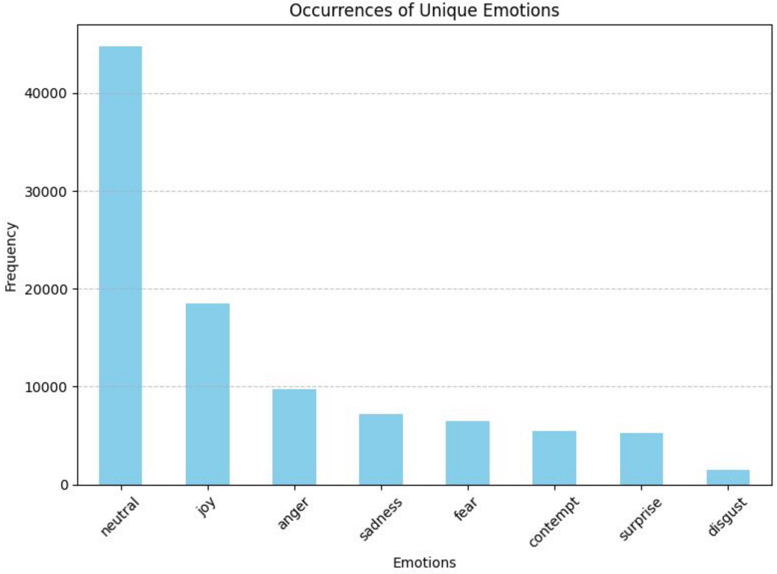
No. of Sentences: 1354



Analysis of Data for Task 2
(EFR in Hindi-English mixed)

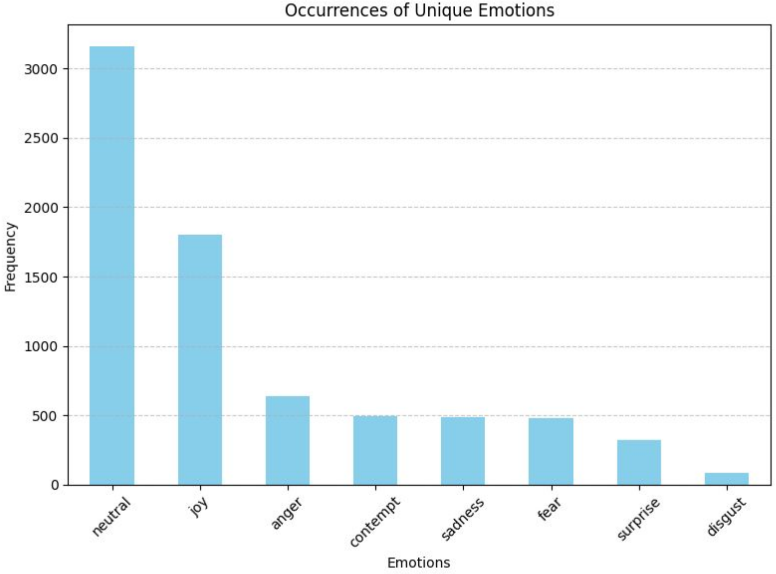
Train Data: No. of Episodes: 4,893

Sentences: 98,777| Triggers: 6539



Val Data: No. of Episodes: 389

Sentences: 7642 | Triggers: 431



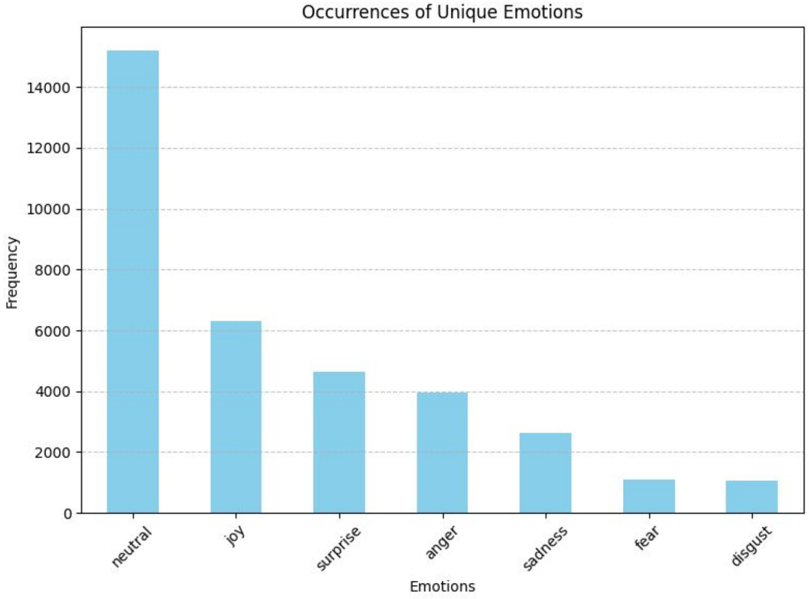
Flip of Emotion one to another in Task 2 Train data

	Neutral	Anger	Surprise	Fear	Joy	Contempt	Sadness	Disgust
Neutral	0	4464	2554	2555	7170	2700	3447	600
Anger	4675	0	411	887	934	466	688	161
Surprise	2253	328	0	550	816	156	346	95
Fear	2751	667	405	0	842	214	306	140
Joy	6645	1275	713	884	0	954	886	221
Contempt	2502	568	197	130	899	0	353	27
Sadness	3482	663	447	315	611	395	0	162
Disgust	580	188	55	75	203	153	93	0

Analysis of Data for Task 3 (EFR in English)

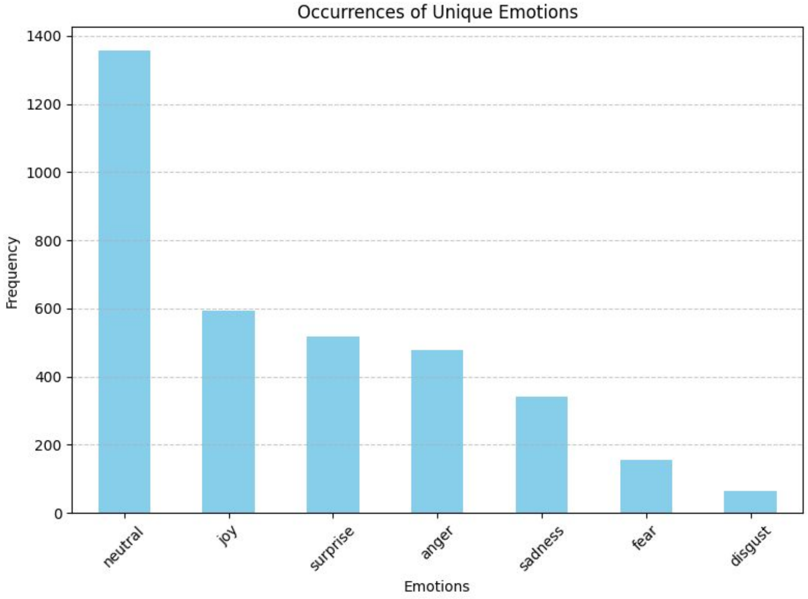
Train Data: Ep. No: 821

Sentences: 34860 | Triggers: 5516



Val Data: Ep. No: 94

Sentences: 3512 | Triggers: 489



Flip of Emotion one to another in Task 3 Train data

	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger
Neutral	0	1866	361	805	2260	357	1255
Surprise	1706	0	105	329	617	143	458
Fear	408	100	0	75	98	16	127
Sadness	685	350	91	0	253	88	285
Joy	2187	628	153	301	0	128	424
Disgust	300	132	39	96	63	0	153
Anger	1200	388	144	283	401	99	0

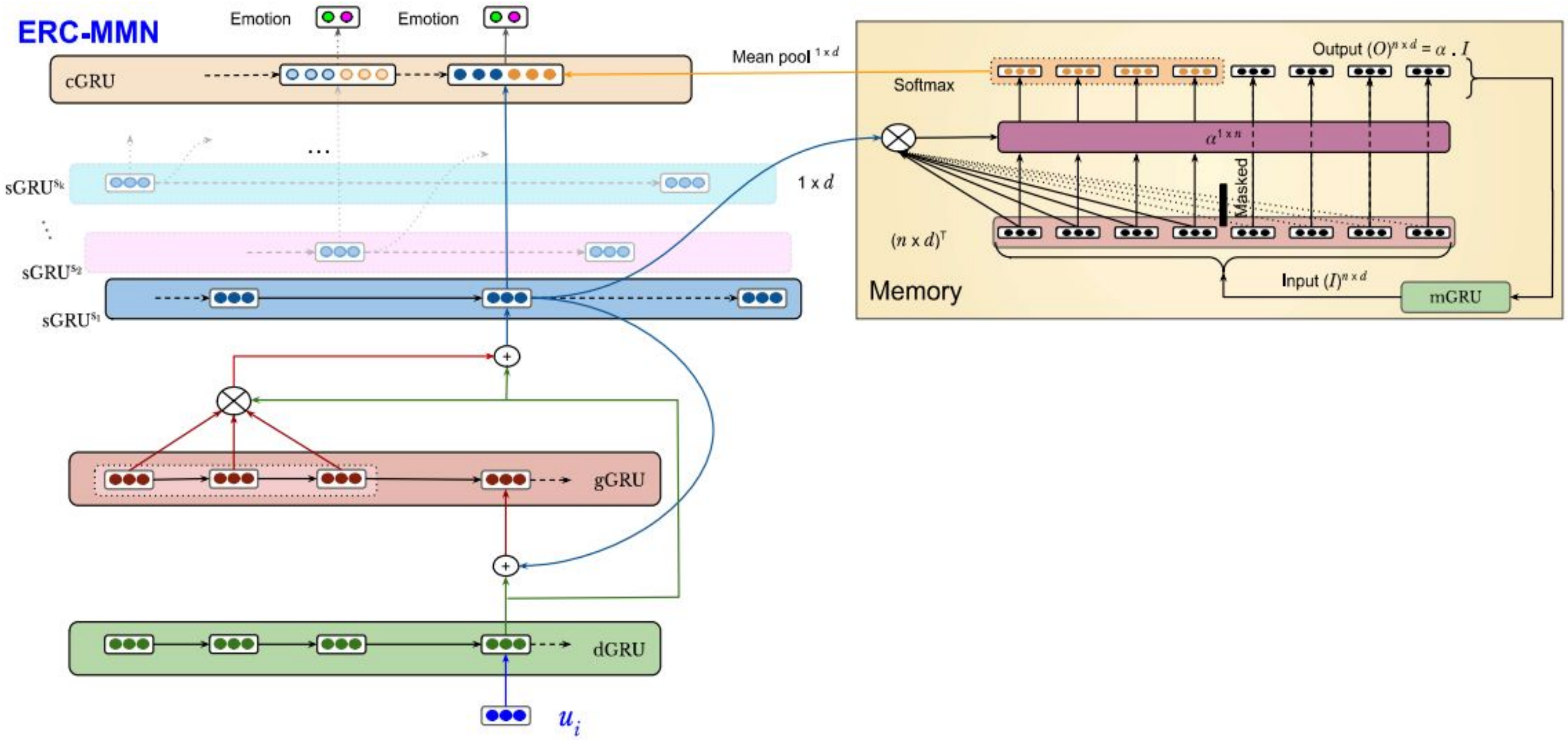
Approaches

1. **Task1 ERC MASAC (HINGLISH)**
   a) ERC-MMN
   b) SPCL-SimCSE
   c) SPCL-HingBERT
   d)SPCL-HingBERT-GRU

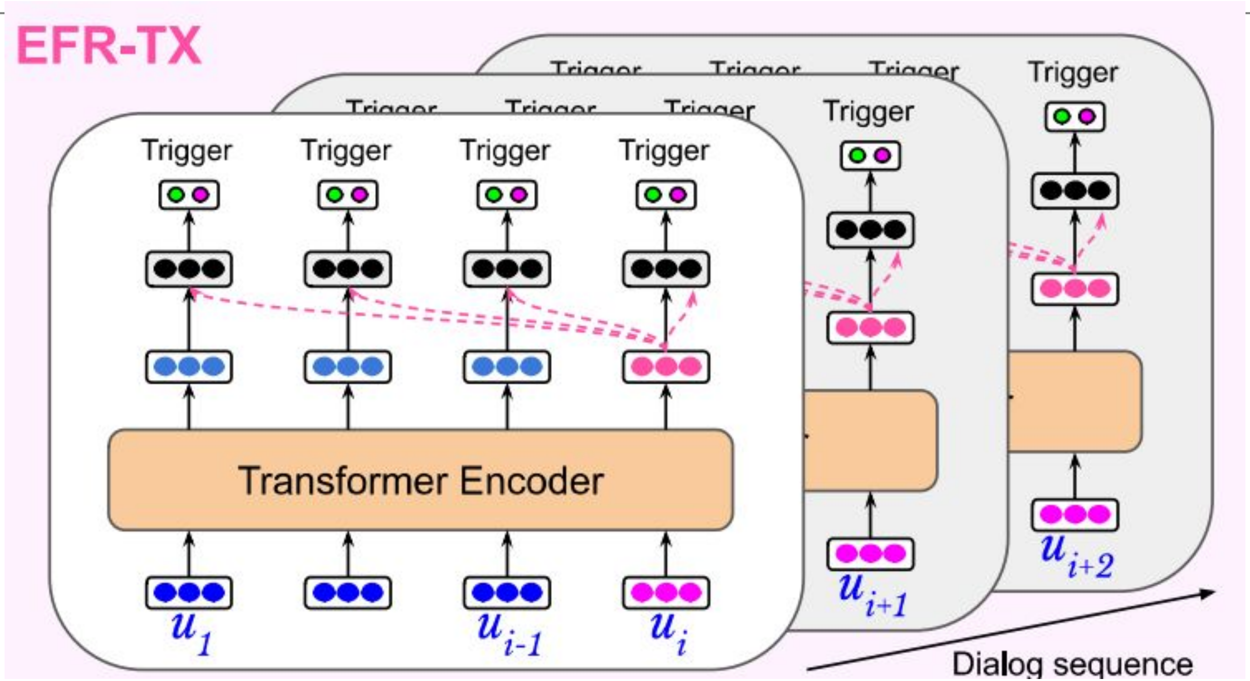
2. **Task2 EFR MASAC (HINGLISH)**
   a)EFR-TX
   b)EFR-TX HYPO

3. **Task3 EFR MELD (ENGLISH)**
   a)EFR-TX
   b)EFR-TX HYPO
   c)SPCL-SimCSE

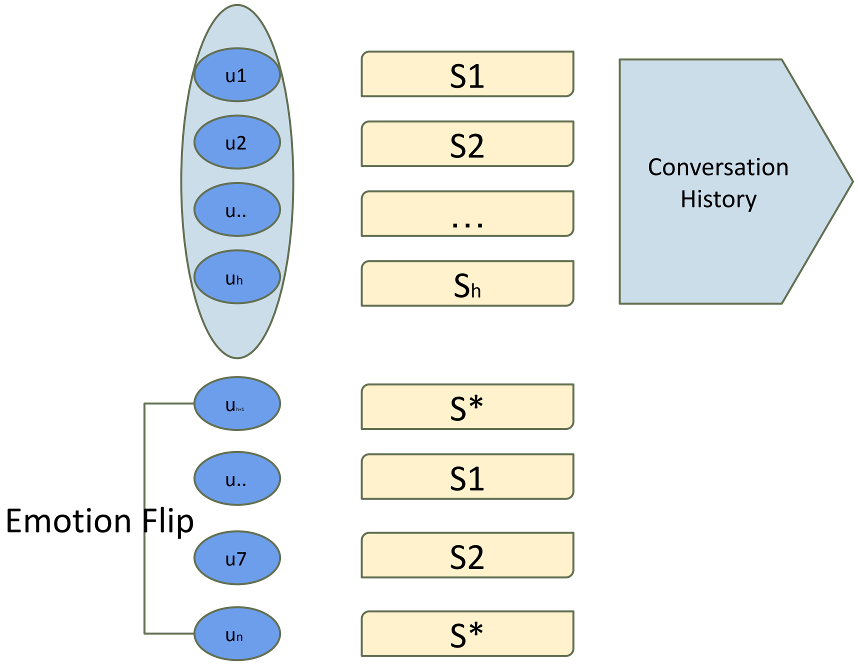
Methodology for ERC-MMN



Methodology for EFR-TX



Our Hypothesis for Emotion Flip



- Highly imbalanced data, Biased towards ‘0’
- Hypothesis: Triggers occur between the emotion flip of $S*$
- We focus on the utterances between $S*$ to detect triggers (including utterances of $S*$)

Impact of Hypothesis on Data

MaSaC:

Dataset	Triggers	Non-Triggers	Average length
MaSaC	6970	99252	19.68
MaSaC (Hypo)	6304	16284	4.25

MELD:

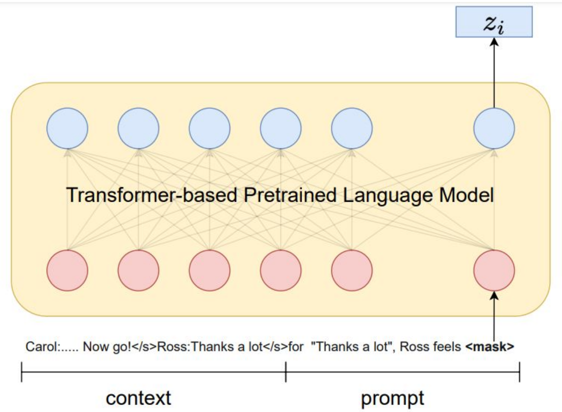
Dataset	Triggers	Non-Triggers	Average length
MELD	6005	32357	8.5
MELD (Hypo)	5330	9735	3.32

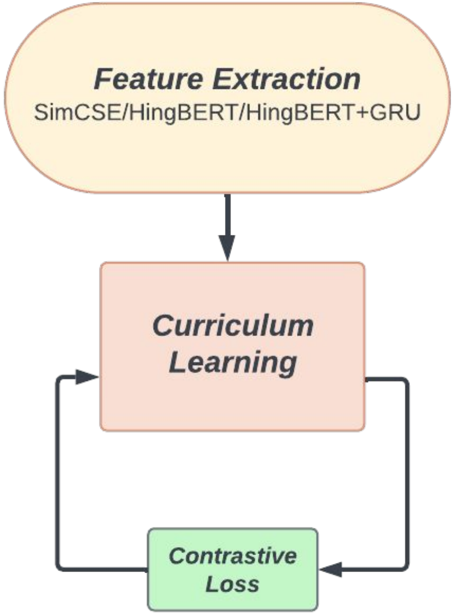
Supervised Prototypical Contrastive Learning

Feature Extraction:

Built a prompt-based context encoder using SimCSE to model the context.

We also experiment with variations of HingBERT To get representations of the MaSaC Hinglish dataset

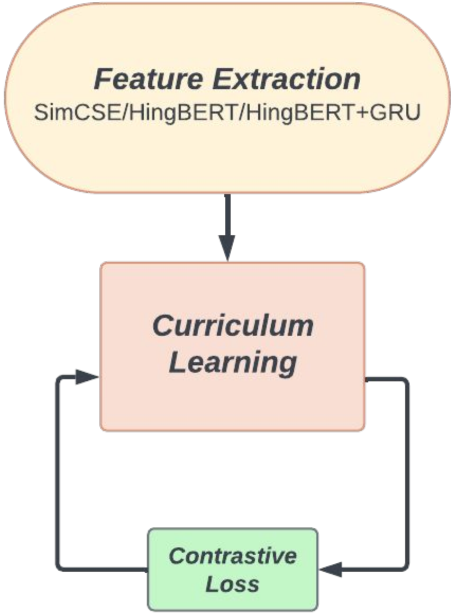




Supervised Prototypical Contrastive Learning

Prototypical Contrastive Loss based Learning:
This loss tackles class imbalance in contrastive learning, using prototype vectors obtained by randomly selecting and averaging samples from a fixed-size representation queue for each emotion category.

Curriculum Learning:
Curriculum learning mitigates the problem by calculating the center of each label class by taking average. Difficulty is determined by the distance of each data point from its label class's center. We use a sampling-based approach to create subsets, ranging from easy to hard, instead of directly splitting the training set.



Experiments

CNN Model for ERC:

Created a CNN based model for Emotion Recognition without taking conversation history into context

ERC MMN and EFR TX:

Modified the ERC-MMN and EFR-TX pipelines to benchmark on EDiReF dataset

EFR-TX with Hypothesis:

The data is highly imbalanced towards 0. So as discussed in previous slides we used our hypothesis to create the new data and kept the EFR-TX model same as before

> **Embeddings for MaSaC:** HingBert, a pre-trained BERT model trained on the Code-Mixed Hindi-English Corpus, to generate word embeddings. We generate sentence embeddings by averaging the individual word embeddings in a sentence.
>
> **Embeddings for MELD**: Sentence Embedding generated using SBert
>
> **Weights:** We set the weights for the Cross-Entropy Loss inversely proportional to the support of the dataset for each emotion

Experiments

SPCL: To extract emotion/trigger representations we use the following methods:

1. **SimCSE-** We feed the training data into SimCSE and then get the last hidden state and then we use the embedding of the special token <mask> as our representation

   Prompt for ERC P_t = for u_t , s_t feels <mask>

   Proposed Prompt for EFR P_t = for u_t , s_t feels e_t is possibly <mask>

   e_t being the emotion of speaker s_t

2. **HingBERT-** To get representations of the MaSaC Hinglish dataset we use the pre-trained HingBERT model to generate utterance representations corresponding to each emotion.

3. **HingBERT with GRU-** GRU-based representation, where we pass individual word embeddings through a GRU to generate sentence representations for each emotion.

Our Results for ERC

Model	F1 score	Accuracy	Precision	Recall	F1 Macro Avg
SPCL SimCSE	**0.45**	0.47	0.45	0.47	**0.33**
ERC MMN Hinglish	0.33	0.46	0.32	0.46	0.11
SPCL HingBERT	0.29	0.42	0.23	0.42	0.09
SPCL HingBERT GRU	0.30	0.47	0.22	0.47	0.08

Our Results for EFR

1. **MaSaC:**

Model	F1 score	Accuracy	Precision	Recall	F1 Macro Avg
EFR-TX MASAC	0.62	0.58	0.69	0.58	0.52
EFR-TX MASAC Hypo	**0.64**	0.63	0.64	0.63	**0.57**

2. **MELD:**

Model	F1 score	Accuracy	Precision	Recall	F1 Macro Avg
EFR-TX MELD	0.60	0.56	0.68	0.56	0.53
EFR-TX MELD Hypo	0.52	0.51	0.58	0.51	0.51
SPCL MELD	**0.80**	0.82	0.79	0.82	**0.56**

Result of Group 9

Task 1:

Model	Weighted F1
Bard	—
ChatGPT 3.5	0.35
ChatGPT 4.0	0.50
ERC-MMN - unweighted loss	0.27
ERC-MMN - weighted loss	**0.92**

Task 2:

Approach	Weighted F1
One Hot Speaker Representation	**0.62**

Task 3:

Models	Weighted F1
EFR-TX	0.66
EFR-TX (Frozen)	**0.70**
One Hot Personality	0.70

Task 3:

Models	Weighted F1
One Hot Personality (Frozen)	0.65
100 Dim Personality	0.66
100 Dim Personality (Frozen)	0.55
One Hot Personality Linear Decoder	0.68
One Hot Personality MLP Decoder	0.67

Result of Group 2

Model	F1 score	Accuracy	Precision	Recall	F1 Macro Avg
Task 1 (ERC)					
Baseline & Proposed	**0.37**	0.53	0.28	0.53	**0.09**
Task 2 (EFR MaSaC)					
Baseline	0.46	0.43	0.67	0.43	0.42
Proposed	**0.54**	0.49	0.67	0.49	**0.46**
Task 3 (EFR MELD)					
Baseline	0.36	0.36	0.61	0.36	0.36
Proposed	**0.38**	0.38	0.62	0.38	**0.38**

Group	**F1 score**	**Accuracy**	**Precision**	**Recall**	**F1 Macro Avg**
Task1					
Our Group	**0.45**	0.47	0.45	0.47	**0.33**
Group 2	0.37	0.53	0.28	0.53	0.09
Group 9	**0.92**	—-	—-	—-	—-
Task2					
Our Group	**0.64**	0.63	0.64	0.63	**0.57**
Group 2	0.54	0.49	0.67	0.49	0.46
Group 9	0.62	—-	—-	—-	—-
Task3					
Our Group	**0.80**	0.82	0.79	0.82	**0.56**
Group 2	0.38	0.38	0.62	0.38	0.38
Group 9	0.70	—-	—-	—-	—-

Comparison of Best Results

Error Analysis

- SPCL was employed to address the challenge posed by an imbalanced dataset. This imbalance arose due to the prevalence of neutral sentences in ERC and the majority of speakers not being the trigger in EFR.

- We encountered challenges, particularly in the generation of sentence representations using HingBERT, as averaging all word embeddings to create sentence embeddings did not yield satisfactory results.

- In an attempt to address the above issue, we experimented with generating sentence representations using a GRU model on individual word embeddings. However, this approach only resulted in minor improvements.

Minor Mistakes in the report

The scores emphasized in the image, as referenced in the report, correspond to the F1 score for class 1, namely the trigger class.

Model	F1 Score	Accuracy	Precision	Recall	F1 Macro Average
EFR-TX MASAC	0.36	0.58	0.69	0.59	0.52
EFR-TX MASAC Hypo	0.41	0.63	0.64	0.63	0.57

Table 3: Results for EFR MaSaC

Model	F1 Score	Accuracy	Precision	Recall	F1 Macro Average
EFR-TX MELD	0.43	0.56	0.68	0.56	0.53
EFR-TX MELD Hypo	0.49	0.51	0.58	0.51	0.51
SPCL MELD	0.80	0.86	0.79	0.82	0.56

Table 4: Results for EFR MELD

ground up, transfer learning is also a viable option to consider.

5. The use of ensembling is a potential strategy since we employed multiple models.

10 Individual Contributions

1. Sachin: Literature Review, CNN based Model, SPCL, Hypothesis Model
2. Yashwant: Literature Review, CNN based Model, SPCL, Hypothesis Model
3. Ritam: Literature Review, Data Analysis, Data preprocessing, Benchmarking on ERC MMN and EFR TX

11 Conclusion

In conclusion we conducted a comprehensive analysis of the dataset. We developed a naive CNN-based model without taking conversation history into account We then benchmarked MaSaC ERC data on ERC-MMN and MaSaC EFR ,MELD EFR on EFR-TX model,we used a pre trained BERT model trained on Code-Mixed Hindi-English Corpus, to generate word embeddings for MaSaC ERC and MaSac EFR in the above models.We then

Contribution of Each Member

1. Sachin: Literature Review, CNN based Model, SPCL, Hypothesis Model

2. Yashwant: Literature Review, CNN based Model, SPCL, Hypothesis Model

3. Ritam: Literature Review, Data Analysis, Data preprocessing, Benchmarking on ERC MMN and EFR TX

Citations

- **InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-task LLMs Framework** - *Shanglin Lei, Guanting Dong, XiaoPing Wang, Keheng Wang, Sirui Wang*
- **Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation** - *Xiaohui Song, Longtao Huang, Hui Xue, Songlin Hu*
- **Hierarchical Dialogue Understanding with Special Tokens and Turn-level Attention** - *Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, Yang You*
- **M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation** - *Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, Naoyuki Onoe*
- **CFN-ESA: A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition** - *Jiang Li, Yingjian Liu, XiaoPing Wang, Zhigang Zeng*
- **Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations** - *Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, Songlin Hu*
- **Discovering Emotion and Reasoning its Flip in Multi-Party Conversations using Masked Memory Network and Transformer** - *Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, Tanmoy Chakraborty*
- **DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation -** *Deepanway Ghosal , Navonil Majumder , Soujanya Poria , Niyati Chhaya and Alexander Gelbukh*
- **ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection -** *Devamanyu Hazarika , Soujanya Poria , Rada Mihalcea , Erik Cambria and Roger Zimmermann*
- **Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos -** *Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, Roger Zimmermann*
- **Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network** - *Wenxiang Jiao,1 Michael R. Lyu,1 Irwin King*

- **Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation** - *Xiaohui Song, Longtao Huang, Hui Xue, Songlin Hu*

Thank you