IITK at SemEval-2024 Task 10: Who is the speaker? Improving Emotion Recognition and Flip Reasoning in Conversations via Speaker Embeddings

Shubham Patel[*]   **Divyaksh Shukla**[*]   **Ashutosh Modi**

Indian Institute of Technology Kanpur (IIT Kanpur)

{devang21, divyakshs23}@iitk.ac.in

{ashutoshm}@cse.iitk.ac.in

Abstract

This paper presents our approach for the SemEval-2024 Task 10: Emotion Discovery and Reasoning its Flip in Conversations. For the Emotion Recognition in Conversations (ERC) task, we utilize a masked-memory network along with speaker participation. We propose a transformer-based speaker-centric model for the Emotion Flip Reasoning (EFR) task. We also introduce *Probable Trigger Zone*, a region of the conversation that is more likely to contain the utterances causing the emotion to flip. For sub-task 3, the proposed approach achieves a 5.9 (F1 score) improvement over the task baseline. The ablation study results highlight the significance of various design choices in the proposed method.

1 Introduction

Conversations between participants carry information that evokes emotions. Emotions include personality, character, temper, and inspiration as the primary psychological parameters that drive them (P S and G S, 2017). Analyzing emotions through language helps uncover the interpersonal sentiments in a conversation at a finer level. This can help build better affective generative models (Goswamy et al., 2020), like chatbots that understand emotion and respond according to a person's behaviors and personality (Kumar et al., 2021; Colombo et al., 2019).

The SemEval-2024 Task 10 (Kumar et al., 2024) aims at Emotion Recognition (ERC), sub-task 1, and Emotion Flip Reasoning (EFR), sub-tasks 2 and 3, in conversations for two languages, namely English and Hindi-English Code-Mixed. ERC refers to identifying the emotion of different utterances. EFR is about identifying those utterances in the dialogue that caused the emotion of a speaker to change.

We build upon the models presented in Kumar et al. (2021) for ERC and EFR. A speaker's personality is likely to influence the emotions developed in other participants (Hazarika et al., 2018a). This inspired us to include information regarding speaker participation to improve the analysis of the emotion of an utterance in conversations. Additionally, for Emotion Flip Reasoning, we propose the Probable Trigger Zone (PTZ), a region of the conversation more likely to consist of the utterance that caused an emotional change in the target participant. This helps us filter out significant non-trigger utterances, reducing the skew in the data. We utilize pre-trained models for computing text embeddings to obtain better representations of utterances.

In sub-task 1, we achieved a weighted F1 score of 45 and 9th rank. For sub-tasks 2 and 3, we secured 5th and 10th position with F1 scores of 56 and 60, respectively. The top scores for each sub-task were 78, 79, and 79, respectively. For sub-task 3, our model improves 5.9 F1 over the baseline model presented in Kumar et al. (2021). The proposed changes have assisted in improving the performance of the system. A limitation of our model is knowing speakers. It might not be possible in all circumstances that this information is available. Also, despite trying to reduce the skew in the data, our model's performance was still impacted. Our models and code can be found here.[1]

2 Related Work

2.1 ERC

The task of emotion prediction has been of active interest in recent years (Witon et al., 2018; Kumar et al., 2020; Keswani et al., 2020; Singh et al., 2021b, 2023, 2021a), including the development of

---

[*] Equal Contributions

[1] https://github.com/Exploration-Lab/IITK-SemEval-2024-Task-10-Emotion-Flip

models like ICON (Hazarika et al., 2018a), COG-MEN (Joshi et al., 2022), Instruct-ERC (Lei et al., 2023) and the models by Kumar et al. (2021). Also, there has been active research in affective text generation (Goswamy et al., 2020). Several datasets exist (Bedi et al., 2023; Poria et al., 2019; Busso et al., 2008) that use one or more additional emotions along with Ekman's scheme (Ekman, 1992) of emotion representation via six emotion classes, namely, *fear*, *anger*, *joy*, *sad*, *disgust*, and *surprise*.

Hazarika et al. (2018a) and Li et al. (2020) highlight the importance of inter and intra-speaker interactions in a conversation. Li et al. (2020) achieves this by using three separate transformer-encoder blocks: (1) Conventional masking: masked multi-head self-attention, (2) Intra-speaker masking: all utterances from other speakers are masked, and (3) Inter-speaker masking: all utterances from the current speaker are masked. While this captures relationships, it does not capture the speaker's personality or presence. Hazarika et al. (2018a) also considers speakers, but it was modeled on the IEMOCAP dataset (Busso et al., 2008) that contains only two participants.

Shapes of Emotion (Bansal et al., 2022), ICON (Hazarika et al., 2018a) and its derived model ERC_MMN (Kumar et al., 2021) proposed the concept of speaker-level outputs, which means that during conversational flow, there is a speaker-level GRU to encode the currently spoken utterance. They achieve this by storing vectors representing each speaker and updating them using the speaker-level GRUs' hidden outputs, which are initialized to 0 during the start of a dialogue.

COGMEN (Joshi et al., 2022) introduces the concept of graphs to conversation flow for emotion recognition. They represent a graph in which each utterance is a node and is related to past or future utterances of the same or different speaker within a time window. CORECT (Nguyen et al., 2023) leverages on COGMEN and introduces speaker embeddings from MMGCN (Wei et al., 2019) to encode each speaker in the conversation for graph-based interaction and pairwise cross-modal feature interaction.

2.2 EFR

Kumar et al. (2021) introduces the relatively new Emotion-Flip Reasoning (EFR) task, which aims to identify past utterances in a conversation that have triggered one's emotional state to flip at a certain time. The task of Emotion-Cause Pair Extraction

(ECPE) (Xia and Ding, 2019) and Emotion Cause Extraction (ECE) (Gui et al., 2016) are similar to EFR, but they aim to extract the causes of emotions from a given text instead of conversations. Kumar et al. (2021) present a transformer-based model for EFR and also measure the performance of baseline models CMN (Hazarika et al., 2018b), ICON (Hazarika et al., 2018a), DGCN (Ghosal et al., 2019), AGHMN (Jiao et al., 2019), and Pointer Network (Vinyals et al., 2017).

2.3 Embeddings

The performance of models on tasks is influenced by the quality of text representation it uses (Asudani et al., 2023). Nayak and Joshi (2022) release HingBERT, a BERT model that has been fine-tuned on Hindi-English Code-Mixed corpus. Muennighoff et al. (2023) introduce the Massive Text Embedding Benchmark (MTEB), which evaluates the performance of text embeddings through different tasks across several datasets. One of the top performers, the `voyage-embeddings`[2], utilize neural-net models to encode the text into text embeddings.

3 Task

SemEval-2024 Task 10: "**E**motion **Di**scovery and its **Re**asoning it **F**lip in Conversations" (Kumar et al., 2024), EDiReF, consisted of three sub-tasks:

1. ERC in Hindi-English Code-Mixed.

2. EFR in Hindi-English Code-Mixed.

3. EFR in English.

Emotion Recognition in Conversations (ERC) is classifying the utterances in a dialogue into one of the given emotion categories. An emotion flip is said to have occurred when a speaker's utterance differs from his/her previous utterance's emotion. Emotion Flip Reasoning (EFR) refers to identifying the utterances (triggers) that caused an emotional flip. These utterances could have been spoken either by the speaker himself or someone else. For the task of ERC, given the utterances in the conversation and corresponding speaker names, the emotion label for each utterance has to be predicted. For the task of EFR, the emotion labels of utterances have also been provided, and the triggers for a given emotion flip have to be predicted.

---

[2]https://docs.voyageai.com/embeddings/

Sub-task	1	2	3
Emotions	8	8	7
Episodes	343	452	833
Utterances	8506	11260	8747
Triggers	-	6542	5575

Table 1: Statistics of the Training Dataset.

Sub-Tasks 1 and 2 tailor the Hindi-English Code-Mixed dataset - MaSaC (Bedi et al., 2023). The training dataset consists of utterances in Roman script (e.g., "*yah plastic ke stickers tumne kahan se khariden?*"). Sub-Task 3 uses the MELD-FR dataset presented in Kumar et al. (2021) built upon the MELD dataset (Poria et al., 2019). Table 1 highlights the overall statistics regarding the training set for each sub-task. Figure 1a and Figure 1b show the distribution of the triggers as a function of the distance from the target utterance for the sub-task 2 and sub-task 3 training datasets, respectively.

4 System overview

Inspired by the use of memory networks by Hazarika et al. (2018a) and Kumar et al. (2021) for emotion recognition, in 4.2 we present our model for the task of ERC, sub-task 1. Inspired by a transformer-based (Vaswani et al., 2017) approach for emotion flip reasoning presented by Kumar et al. (2021), in 4.3 we present our models for sub-tasks 2 and 3.

4.1 Utterance Embeddings

We utilize pre-trained models to compute representations of the utterances in the conversation. Sub-tasks 1 and 2 required the computation of utterance embeddings for code-mixed Hindi-English sentences. We utilized HingBERT to compute the utterance embedding as an average of all the token embeddings in an utterance. Sub-Task 3 consists of utterances in English. We referred to the Massive Text Embedding Benchmark to determine an efficient method to compute utterance embeddings. We experimented with the embeddings presented in Li and Li (2023) and voyage-embeddings, out of which the latter performed better. Hence, we used the `voyage-lite-02-instruct` model with `query_type` as a `document`.

4.2 ERC

We take inspiration from the Masked Memory Network architecture presented by Kumar et al. (2021)
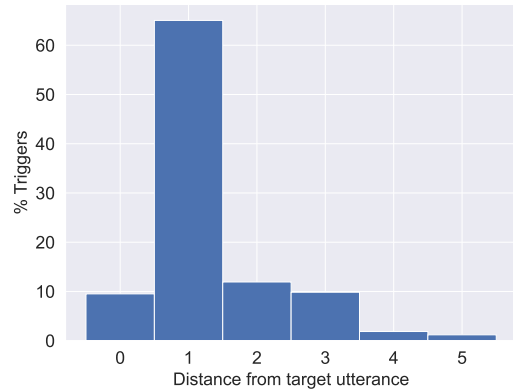
(a) Sub-Task 2.
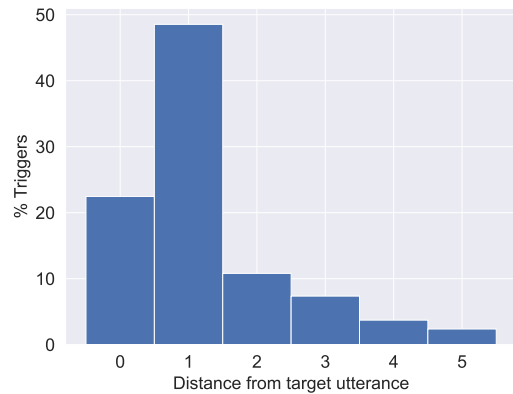
(b) Sub-Task 3.

Figure 1: Distribution of the distance between the target utterance and the causal utterance for emotion flip.

and speaker-specific GRUs proposed by Kumar et al. (2021) and Hazarika et al. (2018a). We used HingBERT to encode each utterance and then pass them through a dialog-level GRU and a speaker-level GRU. The vectors from the global-level GRUs are passed through a memory network through multiple hops (a cycle of reading from memory and writing back to memory is called a hop. The output is taken from the final memory read operation.) Then, attention is computed between the memory and speaker-level outputs while masking future utterances and concatenating with speaker-level outputs to compute conversation-level outputs. Finally, the obtained features are passed through a trainable linear layer for predicting emotion class. Figure 2 shows the model architecture.

Notations: u_t denotes the embedded utterance at time t in a dialogue, while $s_t^{(k)}$ denotes the kth speaker embedding for utterance u_t. $attention(q, k, v)$ is the attention operator applied on query q, key k and value v. r is the number of
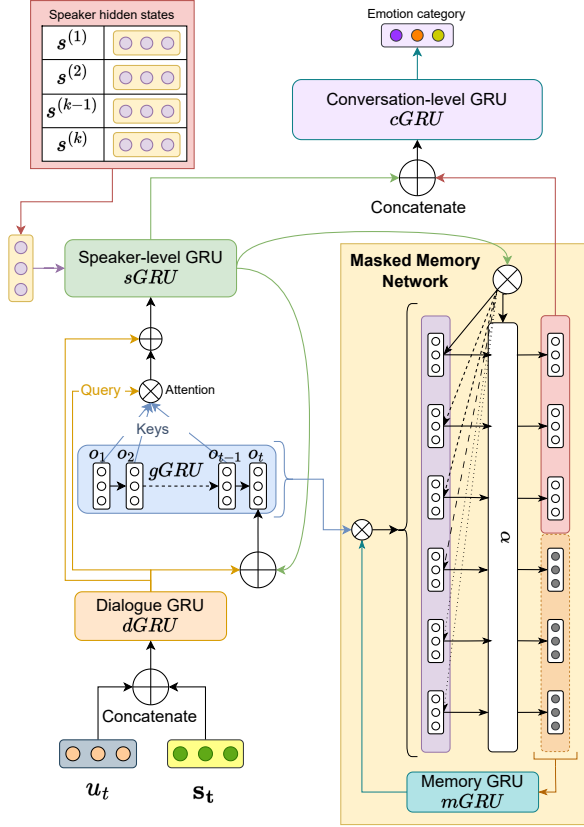
Figure 2: Masked Memory Network with Speaker-Embeddings concatenated with utterance embeddings. Speaker-embeddings are one-hot vectors of 6-dimensions which store 1 at the index of the top-6 speakers, otherwise 0.

hops in the memory layer, inspired from Hazarika et al. (2018a).

Dialogue-level GRU $dGRU$**:** This recurrent unit gives a dialogue-level representation of the u_t and gives output as do_t.

$$do_t = dGRU(u_t \oplus s_{kt})$$

Global-level GRU $gGRU$**:** This recurrent unit gives a global-level representation of the utterances $u_{(1:t)}$ till time step t, as $o_{(1:t)}$.

$$o_t = gGRU(do_t \oplus so_{t-1})$$

Attention Module: Attention is computed between do_t as query and value and $o_{(1:t-1)}$ as keys to obtain attention-based context for speaker-level GRU layer.

$$attention(do_t, o_{(1:t-1)}, do_t)$$

Speaker-level GRU $sGRU$**:** This gives a speaker-level recurrent unit that takes inputs $attention$ and speaker hidden state $s^{(k)}$ (taken from a dictionary of size k) and gives outputs so_t. The hidden output

replaces the dictionary entry for $s^{(k)}$. At the start of a dialogue, the dictionary is empty, and the default hidden state for a new speaker is a zero vector.

$$so_t = sGRU(attention(do_t, o_{(1:t-1)}, do_t) + do_t)$$

Masked-Memory Attention: A memory vector, which represents the previous dialogues and utterances, is obtained by passing $o_{(1:t)}$ through a memory GRU ($mGRU$). This then goes through masked attention with the so_t while masking future utterances and a softmax activation α to give attention weights to each utterance in $o_{(1:t-1)}$. This is then used to update the memory vectors via the $mGRU$ and is concatenated with so_t as an input to $cGRU$.

$$temp = mGRU(o_{(1:t)})$$
$$mem^r = masked\_attention(temp, so_t)$$
$$temp = mGRU(mem^{r-1})$$

Conversation-level GRU $cGRU$**:** This layer represents the conversation flow of the dialogue and takes inputs as the concatenation of so_t and masked attention output, to give conversation-level features.

$$co_t = cGRU(mem^r + so_t)$$

Finally, the outputs of the $cGRU$ are used to compute the emotion class.

$$e_t = W.co_t + b$$

4.3 EFR

4.3.1 Baseline

In Kumar et al. (2021), the authors propose a transformer-based model for the task of EFR, whose architecture is as follows. Utterance embeddings for each utterance are computed using BERT (Devlin et al., 2019). The utterance embeddings of a conversation are passed through a transformer to compute a contextualized utterance embedding for each utterance. The emotion classes for each utterance are encoded as a one-hot vector and passed through a GRU to compute the emotion-history vector. For each utterance, its contextualized embedding, the contextualized embedding of the target utterance, and the emotion-history vector are passed through a linear layer to make a prediction.

4.3.2 Speaker-Aware Embeddings

As highlighted by Li et al. (2020) and Hazarika et al. (2018a), speaker interaction also drives the emotion of an utterance. Unlike their approach for modeling intra and inter-speaker interaction, we believe that the participation of certain speakers in the conversation drives the flip in the emotion of an utterance. Providing information regarding the speaker will help the model learn the nature of the specific speakers. To incorporate this, we utilize that the speakers in the test and the train set overlap.

An aspect of the conversation that has not been captured by the baseline model regarding the speakers of an utterance. In the baseline model, each utterance is treated as an independent text, and its embedding has been computed. This has failed to incorporate the information regarding who the speaker of a given utterance was. To incorporate this aspect, we concatenate the utterance embeddings with a one-hot vector denoting the speaker to create speaker-aware embeddings. This equips the model with the ability to capture the behavioral trends of specific speakers.

4.3.3 Probable Trigger Zone (PTZ)

We propose a hypothesis regarding the possible location of triggers. We divide the conversation into two parts. The first part consists of the utterances before the target speaker's previous utterance. The second part consists of utterances from the target speaker's previous utterance to his target utterance. We call the second part the Probable Trigger Zone (PTZ).

We hypothesize that no triggers lie in the first part of the conversation. Since the target speaker's emotion has flipped during the second part of the conversation, it is more likely that the causes for the emotion flip lie in the second part. Suppose the trigger causing the target emotion had been in the first part. In that case, it is more likely that it would have already caused the emotion of the previous utterance of the target speaker to flip. Then, the same emotion would have been carried to the target utterance, wrongly implying that no emotion flip occurred at the target utterance. To incorporate the hypothesis, we mask any predictions made by the model outside the Probable Trigger Zone. In section 5, we discuss how PTZ helps to reduce skew in the dataset.

For example, consider the conversation in Figure 3. Here, the target speaker is *Ross* with the target
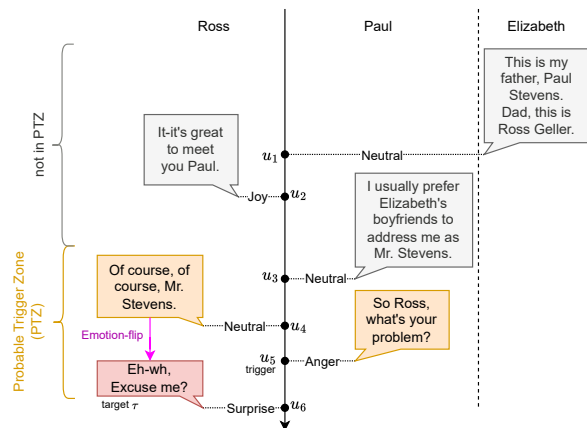

Figure 3: Probable Trigger Zone.

utterance u_6 and his previous utterance u_4. The probable trigger zone consists of utterances from u_4 to u_6. Due to the "surprise-causing" statement u_5 in PTZ, *Ross's* emotion flips from *Neutral* to *Surprise*. If this "surprise-causing" statement had been present outside the PTZ, i.e., before the previous utterance u_4, then the emotion of u_4 would likely have been *Surprise*.

4.3.4 Emotion-Aware Embeddings

Using an Emotion-GRU, the baseline model computes an emotion-history vector from the emotion labels. It uses this emotion-history vector in the final linear layer to predict the utterance label. A possible shortcoming of the above is that the linear layer has access only to the emotion history rather than to the emotion labels of the individual utterances. Also, the transformer layer cannot access the emotion labels while computing the contextualized utterance embeddings. Providing those to the transformer will also allow the embeddings to be emotion-aware. We concatenate our speaker-aware utterance embeddings and one-hot emotion labels to incorporate the above.

4.3.5 Model Functioning

Figure 4 presents the model architecture used for the task of EFR. The target utterance is denoted by the subscript τ. Each utterance u_t of a dialogue d is concatenated with its true emotion label e_t and one-hot speaker embedding s_t. This is then passed through the transformer to take into account the context. The Emotion-GRU computes the emotion-history vector. For each utterance, its and the target utterance's contextualized representation and the emotion-history vector are passed through a linear layer to make the prediction.
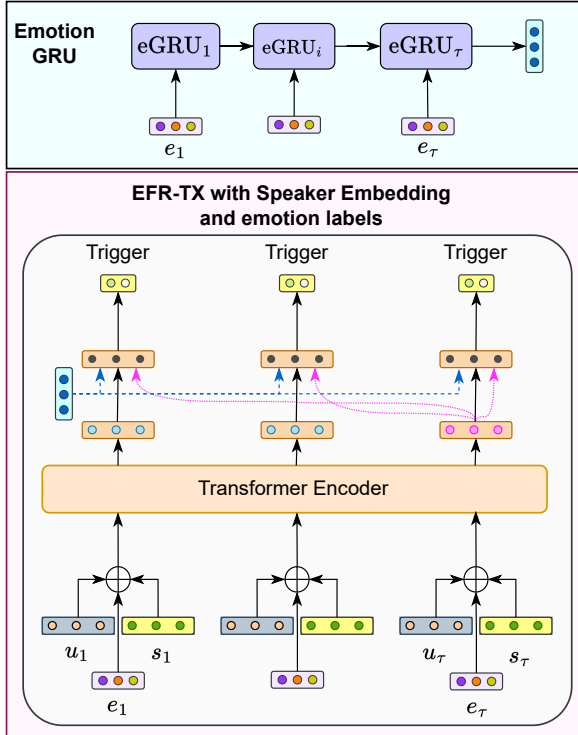
Figure 4: Architecture of the model proposed for the task of Emotion Flip Recognition.

5 Experimental setup

5.1 Training Details

For sub-task 1, we chose a sequence length seq_len of 15, i.e., we break a long conversation into disjoint smaller conversations with utterances less than equal to seq_len. For sub-tasks 2 and 3, we use a window size w of 5. We consider only the last w utterances in a conversation to predict the trigger, i.e., $U_{n-w+1}, U_{n-w+2}...U_{n-1}, U_n$. Table 2 contains details of the hyperparameters we used to train the models. To limit the size of the vector denoting the speaker, keep retained information regarding the top $k = 6$ speakers. We chose the top 6 speakers since this covered nearly $80-85\%$ of utterances in the corpus. We used Adam optimizer (Kingma and Ba, 2017) for all the sub-tasks, with a weight decay of 1e-5. Training of models has been done using Kaggle[3] P100 GPUs.

5.2 Effect of Hypothesis and Sequence Length

In Table 3 and Table 4, we highlight the impact of the hypothesis and selection of sequence length on the datasets. On reducing the window size w to 5, a significant number of negative labels have been eliminated, while there has not been much

---

[3] https://www.kaggle.com/

Sub-Task	1	2	3
Embedding Size	768	768	1024
Batch Size	64	2000	1000
Learning Rate	1e-04	5e-07	5e-07
Weights	Inv Sqrt	Inv	Inv
Epochs	100	1000	1000
Best Epoch	80	299	549
Training Time	10 hr	3 hr	3 hr

Table 2: Hyperparameters for each of the sub-tasks. Weights refers to the weights in the cross entropy loss. *Inv*: Inverse of the supports. *Inv Sqrt*: Inverse of the square root of the supports.

Dataset	0	1	Ratio
Original	92233	6544	14.1
Setting 1	17539	6425	2.7
Setting 2	11535	5839	2.0

Table 3: Effect of PTZ on Dataset, Sub-Task 2. *Setting 1* and *Setting 2* as defined in Section 5.2.

impact on the number of positive labels. Applying the hypothesis has helped mitigate the skew in the data, although there has been a slight impact on the number of positive labels. *Setting 1* refers to considering only the utterances within the window size $w = 5$. *Setting 2* refers to considering utterances that are both within the window and in the probable trigger zone.

Dataset	0	1	Ratio
Original	29416	5575	5.3
Setting 1	13483	5177	2.6
Setting 2	7834	4542	1.7

Table 4: Effect of PTZ on Dataset, Sub-Task 3. *Setting 1* and *Setting 2* as defined in Section 5.2.

6 Results

For Sub-Task 1, the dataset consisted of a non-uniform distribution of labels, with *neutral* being the most frequent. A model that predicts the emotion category of each utterance to be *neutral* achieves a weighted F1 of 24.36. We consider this as a simple *neutral* baseline for sub-task 1. For Sub-Task 2, we have kept the baseline as a rule-based model that predicts the previous utterance to be a trigger and the rest of all utterances non-triggers. The data for the second sub-task is highly skewed

as can be seen in Figure 1a. Due to this baseline performs exceptionally well, as can be observed in Table 5. For Sub-Task 3, we use ERC^{True} EFR-TX from Kumar et al. (2021) as the baseline.

Sub-Task	Model	Metric	Value
2	Rule-Based	F1	79.15
3	ERC^{True} EFR-TX	F1	53.9

Table 5: Baselines for various Sub-Tasks.
Rule-Based: A rule-based model that predicts the previous utterance as a trigger and the rest as non-triggers.

6.1 Model Performance

We have highlighted the performance of our models on the test data in Table 6. For sub-task 2, we get precision and recall scores of 0.73 and 0.83, respectively. For sub-task 3, we get precision and recall scores of 0.74 and 0.80, respectively.

6.2 Error Analysis

For sub-task 1 and sub-task 3, our model performed better than the baselines, but not for sub-task 2. For sub-task 1, the dataset consisted of a non-uniform distribution of labels in the training dataset. Due to this skew in the data, the model has shown different performances for different labels. The predictions for sub-task 1 have been highlighted in Figure 5. For EFR, the usage of a window size $w = 5$ utterances has helped to eliminate a large number of non-triggers. Due to this, the model's predictions have many true negatives, as can be seen in Table 7 and Table 8. But despite this, there was still skew in the data, which impacted the model's performance in predicting the minority class of triggers. The data for sub-task 2 is highly skewed, as can be seen in Figure 1a. We suspect this is why our model has performed poorer than the baseline.

Figure 6 is an example of an erroneous emotion labeling of the model on the test set of sub-task 1 (ERC). Here, the utterance marked in red has the true label as 'Fear,' but the model predicts 'Neutral.' This is due to the sharp change in conversation context at u_5. Also, 'hahhaha' directly corresponds to laughing, but in this case, the speaker at u_6 utters 'hahhaha' as he is worried that the inspector is looking for 'Sharman'. The speaker, Indravardhan, who continuously interacted with the inspector, showed neutral emotion. This led to storing vectors corresponding to neutral for Indravardhan in the memory network, leading to misclassification of emotion at

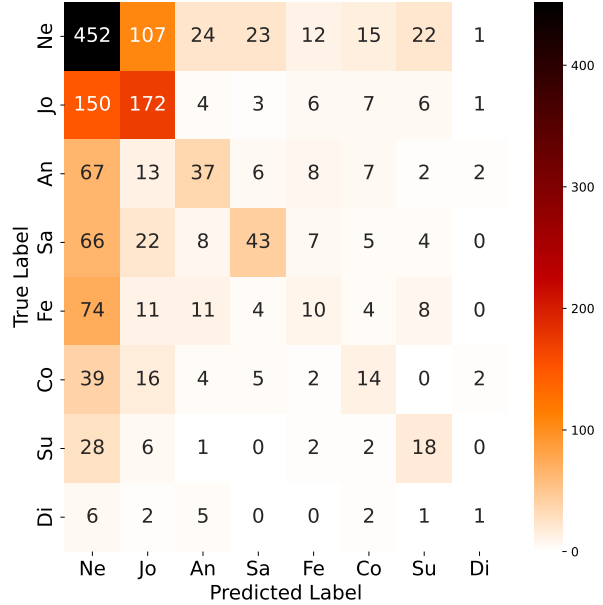| Sub-Task	Metric	Our	Best	Rank
1	Weighted F1	44.80	78	9
2	F1	56.35	79	5
3	F1	59.78	79	10

Table 6: Results on the Test Set.

Figure 5: Confusion Matrix for Sub-Task 1.

u_6.

6.3 Ablation Study

The application of the hypothesis has assisted in removing a few of the wrongly guessed triggers. This has improved the model's performance, as seen in Table 9. We also experimented with another approach of making predictions only inside the PTZ instead of masking the outside labels. This was done by training in the model and making predictions only using the utterances within the probable target zone. In this case, the model's performance was poorer. We suspect this is because the context the model gets in the latter is more restricted than in the first case. Due to this, the model is not able to make predictions effectively.

7 Conclusion

In this paper, we discussed approaches to improve the masked memory network architecture for emotion recognition (ERC) and transformer-based architecture for emotion flip reasoning (EFR) by incorporating speaker information into the embeddings and making better use of the emotion labels. We also employed an approach of focusing the

		Predicted	
		0	**1**
True	**0**	6943	331
	1	123	293

Table 7: Confusion Matrix Sub-Task 2.

		Predicted	
		0	**1**
True	**0**	6735	738
	1	356	813

Table 8: Confusion Matrix Sub-Task 3.

model's prediction on more likely regions to identify triggers by defining the Probable Trigger Zone in conversations. This, along with considering a window of last-few utterances, assisted in reducing the bias in the data.

Limitations: Our model assumes that the training and testing data consist of the same speakers. While this would be true for many benchmark datasets of emotion analysis in conversations, it might not be true in all real-world circumstances. Another limitation of the proposed approach is the training time.

Future Work: In the future, we can include speaker information across dialogues for ERC to capture better speaker semantics by using learnable embeddings for each speaker updated by the hidden outputs of the speaker-level GRU. However, to apply the above, we need to know the number of speakers in the datasets, training, and testing. Additionally, the model becomes dependent on the number of speakers.

A possible approach to help mitigate the assumption of having common speakers and knowing the number of speakers in the training and test time could be exploring further modeling inter and intra-speaker dependencies, as shown in Li et al. (2020) and Hazarika et al. (2018a). They propose models that capture speaker relationships but are not dependent on the number of speakers.

Mitigating the issues of skewed data can be further explored to enhance the system's performance. Also, addressing other aspects of conversations, such as whom the statement is being told to and treating names of speakers in utterances differently from simple pronouns, can be explored.
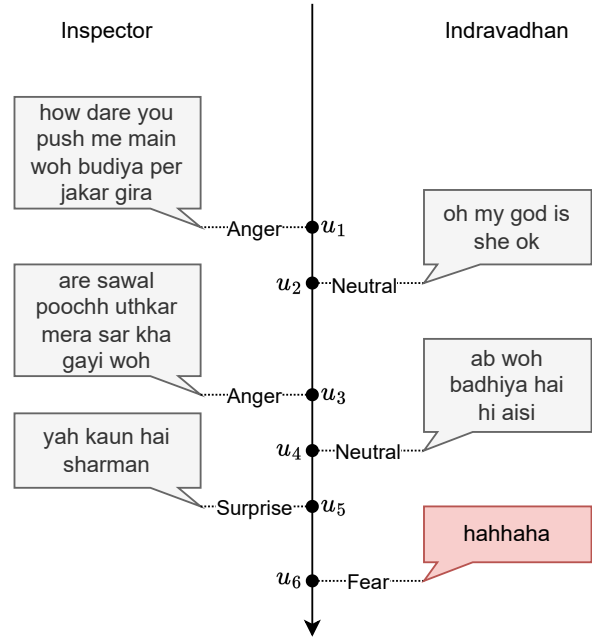

Figure 6: Example of an erroneous emotion labeling from the model. The true label is 'Fear,' but the model predicted 'Neutral.'

Sub-Task	Masks	F1 Before	F1 After	Change
2	1	56.29	56.35	+0.06
3	78	58.68	59.78	+1.10

Table 9: Improvements by PTZ.
Masks: The number of trigger predictions made by the model outside the Probable Trigger Zone, which had been masked to 0.

References

Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. 2023. Impact of word embedding models on text analytics in deep learning environment: a review. *Artif. Intell. Rev.*, 56(9):10345–10425.

Keshav Bansal, Harsh Agarwal, Abhinav Joshi, and Ashutosh Modi. 2022. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 44–56, Virtual. International Conference on 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 Transactions on Affective Computing*, 14(2):1363–1375.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S.

Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resources &. Evaluation*, 42(4):335–359.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. 2020. Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2019. Real-time emotion recognition via attention gated hierarchical memory network.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. 2020. IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of Internet memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1135–1140, Barcelona (online). International Committee for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Ayush Kumar, Harsh Agarwal, Keshav Bansal, and Ashutosh Modi. 2020. BAKSA at SemEval-2020 task 9: Bolstering CNN with self-attention for sentiment analysis of code mixed text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1221–1226, Barcelona (online). International Committee for Computational Linguistics.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. 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, 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. *CoRR*, abs/2103.12360.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Jiangnan Li, Zheng Lin, Peng Fu, Qingyi Si, and Weiping Wang. 2020. A hierarchical transformer with speaker modeling for emotion recognition in conversation.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Cam Van Thi Nguyen, Tuan Mai, Son The, Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

SREEJA P S and Mahalakshmi G S. 2017. Emotion models: A review. *International Journal of Control Theory and Applications*, 10:651–657.

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. Association for Computational Linguistics.

Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021a. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 84–91, Online. Association for Computational Linguistics.

G. Singh, D. Brahma, P. Rai, and A. Modi. 2021b. Fine-grained emotion prediction by modeling emotion definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Los Alamitos, CA, USA. IEEE Computer Society.

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2023. Text-based fine-grained emotion prediction. *IEEE Transactions on Affective Computing*, pages 12–12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2017. Pointer networks.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1437–1445, New York, NY, USA. Association for Computing Machinery.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at IEST 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253, Brussels, Belgium. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.