Emotion Flip Reasoning in Multiparty Conversations

Shivani Kumar, Shubham Dudeja, Md Shad Akhtar, Tanmoy Chakraborty

***Abstract*—In a conversational dialogue, speakers may have different emotional states and their dynamics play an important role in understanding dialogue's emotional discourse. However, simply detecting emotions is not sufficient to entirely comprehend the speaker-specific changes in emotion that occur during a conversation. To understand the emotional dynamics of speakers in an efficient manner, it is imperative to identify the rationale or instigator behind any changes or flips in emotion expressed by the speaker. In this paper, we explore the task called Instigator based Emotion Flip Reasoning (EFR), which aims to identify the instigator behind a speaker's emotion flip within a conversation. For example, an emotion flip from *joy* to *anger* could be caused by an instigator like *threat*. To facilitate this task, we present MELD-I, a dataset that includes ground-truth EFR instigator labels, which are in line with emotional psychology. To evaluate the dataset, we propose a novel neural architecture called TGIF, which leverages Transformer encoders and stacked GRUs to capture the dialogue context, speaker dynamics, and emotion sequence in a conversation. Our evaluation demonstrates state-of-the-art performance ($+4-12\%$ increase in F1-score) against five baselines used for the task. Further, we establish the generalizability of TGIF on an unseen dataset in a zero-shot setting. Additionally, we provide a detailed analysis of the competing models, highlighting the advantages and limitations of our neural architecture.**

***Impact Statement*—Emotions play a pivotal role in deciding the impact of a statement uttered. However, in a conversational setting, simply identifying the emotions of utterances in a dialogue is not enough to characterize the emotional dynamic of the speaker. To this end, the proposed task of emotion-flip reasoning is eminent. The proposed flip explanations via triggers and instigators can help scrutinise how a particular type of remark or expression affects the end listener. A response generation mechanism can use these triggers or instigators as feedback to steer a conversation so that the user feels chipper.**

***Index Terms*—Emotion flip reasoning, instigators, emotion reasoning, conversational dialogues.**

I. Introduction

Understanding emotions is essential for assessing the current state of a speaker in a conversation. Consequently, there has been a significant amount of research in this field [1]. Emotional awareness has proven beneficial in areas that involve aspect analysis of users such as social media [2], [3], [4], and e-commerce [5]. Initial studies focused on standalone texts like tweets [2], [3] to extract the appropriate emotions. However, with the advent of online dialogue agents, the focus of emotion analysis has shifted towards conversation data, usually termed as Emotion Recognition in Conversation (ERC) [6]. Here, the input is a sequence of utterances or a dialogue, instead of isolated texts, and the aim is to identify the emotion of each dialogue utterance. Though emotion is an imperative aspect of a conversation, we posit that it is insufficient to simply identify the speakers' emotion in a dialogue. To reason out the change/flip in emotion of a speaker, a more detailed analysis is required. To this end, we expore the task – **Emotion Flip Reasoning (EFR)** [7].

EFR deals with identifying the cause/reason behind an emotional flip of a speaker in a dialogue. The entire EFR pipeline works in three stages-

(a) Given a sequence of dialogue utterances with emotion labels, the first stage of EFR identifies the utterance where a speaker experienced a flip of emotion.
(b) In the second stage, EFR identifies utterances or triggers responsible for the emotion flip.
(c) Finally, EFR assigns psychologically motivated [8], [9] instigator labels to triggers to explain the emotion flip.

This paper focuses on the third stage, as the first stage can be effortlessly executed from a dialogue with emotion labels, while the second stage [7] is inherent in the third.

Following the cognitive appraisal theory [9], we define a finite set of 27 instigators to reason out flips. Here, we do not account for implicit emotion flips, i.e., emotion flips due to the absence of explicit instigator (e.g., verbal articulation). For instance, emotion flips due to reminiscence can be regarded as implicit. On the other hand, an external trigger is associated with an emotion flip that occurs due to something mentioned in the text (e.g., a person getting scolded).

The task of EFR has the capability to improve user experience in a conversational dialogue system especially in empathetic response generation [10], [11], [12]. A dialogue agent can use EFR triggers for emotion flips in a feedback mechanism. It can award/penalize the agent and the response generator to repeat/avoid using a similar utterance in future conversations. Moreover, the knowledge of instigators can be used *to explain* such emotion flips.

Problem Statement: The EFR task can be formally defined as follows: Given a sequence of n tuples of the form $\langle u_i, s_j, e_k \rangle$ in a multiparty conversation, where $s_j \in S$ is a speaker from a predefined speaker set S, $e_k \in E$ is a set of emotion labels, and $u_i \in D$ is an utterance of the dialogue D, we associate instigator label(s) with an utterance u_l if it causes a flip/change in emotion of a speaker $s_m \in S$ in their consecutive utterances in the conversation. We define the objective in the following three steps.

- First, we identify all utterances u_i in a dialogue which experience an emotion flip of some speaker s_j – when the emotion of the utterance is different from the emotion

28th November 2022. This work was supported by ihub-Anubhuti-iiitd Foundation, set up under the NM-ICPS scheme of the DST.

Shivani Kumar was with Indraprastha Institute of Information Technology Delhi, India. (e-mail: shivaniku@iiitd.ac.in).

Shubham Dudeja was with Indraprastha Institute of Information Technology Delhi, India. (e-mail: shubham19053@iiitd.ac.in).

Md Shad Akhtar was with Indraprastha Institute of Information Technology Delhi, India. (e-mail: shad.akhtar@iiitd.ac.in)

Tanmoy Chakraborty was with Indian Institute of Technology Delhi, India. (e-mail: tanchak@iitd.ac.in)

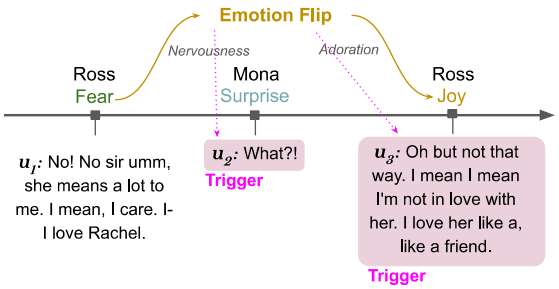


Fig. 1: Example of an emotion flip with triggers and instigators. Ross's emotion flipped from *Fear* (u_1) to *Joy* (u_3) due to two trigger utterances (u_2 and u_3) caused by the instigators, *nervousness* and *adoration*, respectively.

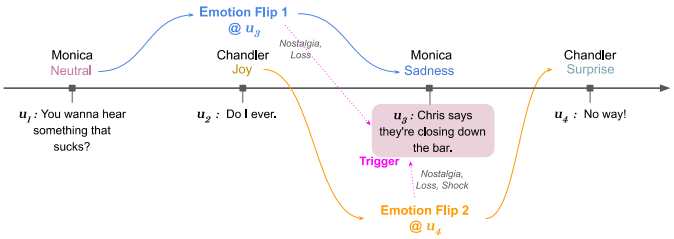


Fig. 2: Example of an emotion flip with self-trigger. Monica's emotion flipped from *Neutral* (u_1) to *Sadness* (u_3) due to one trigger utterance (u_3 itself) caused by the instigators *nostalgia* and *loss*. The other speaker's emotion then flipped from *Joy* (u_2) to *Surprise* (u_4) due to a single trigger utterance (u_3 again) caused by the instigators *nostalgia*, *loss* and *shock*.

associated with the last utterance of speaker s_j. We call u_i as the target utterance.

- Second, for each emotion flip utterance u_i, we mark a set of utterances $u_k \in \{u_1, u_2, \cdots, u_i\}$ which are responsible for the emotion flip at u_i.
- Finally, we associate psychologically motivated instigator labels (c.f. Section III) to each u_k for a target utterance u_i.

Figure 1 illustrates an example of emotion flip with corresponding instigators. It shows a multiparty scenario where three speakers are engaged in a conversation. There are two emotion flips – $\langle u_1, Ross, fear\rangle \rightarrow \langle u_3, Ross, joy\rangle$ and $\langle u_3, Ross, joy\rangle \rightarrow \langle u_5, Ross, anger\rangle$. The first flip occurs due to two trigger utterances, u_2 and u_3, each evoking the feeling of nervousness and adoration in the speaker (Ross). Consequently, the instigator labels for the concerned flip would be *Nervousness* and *Adoration*. On the other hand, the trigger for the second flip is a single utterance (u_4), and the corresponding instigator labels are *Annoyance* and *Challenge* as the trigger instigates the notion of annoyance and challenge in the speaker (Ross). This example highlights the case when more than one trigger utterance can cause an emotion flip.

In Figure 2, we show another example from our dataset. It shows a dyadic conversation having two emotion flips (u_3 and u_4) corresponding to two speakers (Monica and Chandler) involved in the conversation. The emotion flip at u_3 is an example of a self-trigger emotion flip – the responsible utterance (or trigger) is u_3 itself. Moreover, the same utterance u_3 acts as the trigger for both the emotion flips but causes different instigators in the two cases. It is interesting to note that the same utterance causes the emotion *sadness* in one speaker while the emotion *surprise* in another. This highlights the importance of identifying the emotion instigators to understand emotion dynamics completely.

Contributions: We summarise our contributions below-

- We propose a **novel task** – Emotion Flip Reasoning (EFR) in conversational dialogue that aims to explain the shift in emotion for a speaker present in the dialogue.
- We carefully draft a set of ground-truth labels, called **instigators**, to explain an emotion flip.
- We develop a new dataset, called **MELD-I**.
- We design **TGIF**, a Transformer and GRU-based architecture for the EFR task of multi-label instigator classification.

We highlight some related literature in the next section followed by the illustration of the datset used in Section III. Our proposed methodlogy to solve the EFR task is represented in Section IV while the next section shows the experimental setup and results. Finally we analyse the obtained results and conclude in the last section.

The source code of TGIF and the proposed dataset, MELD-I, along with the execution instructions are available at https://github.com/LCS2-IIITD/EFR-Instigators.git.

II. Related Work

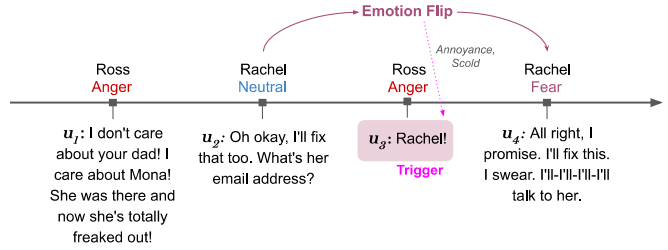
Emotion recognition. Earlier studies in the field of emotion analysis [13], [14], [15], [16], [17], [18] dealt with only standalone inputs. [19] developed a dataset of $5,553$ tweets and manually annotated them with emotion labels. Further, they evaluated this dataset using standard machine learning techniques. Recently [20] used the weighted k-nearest neighbour approach in order to provide an explainable model for emotion detection in tweets. Mulimodality, like speech [21] and visual signals [22], [23], is also a well explored topics in the domain. A detailed survey is provided by [24]. However, these studies are performed for standalone text, which lacks any contextual information.

Emotion and conversation. Recently, the focus of emotion detection has shifted to conversations (emotion recognition in conversation or ERC). It has gained significant popularity due to numerous applications. Existing literature suggests that a wide range of deep learning methods have been applied to address the Emotion Recognition in Conversation (ERC) task [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37]. ICON [25] used a memory network architecture to model the interaction between self and inter-speaker states in two-party conversations. On the other hand, the use of external knowledge was explored in [26] along with a hierarchical self-attention mechanism to detect emotions in conversation. [27] used a party ignorant framework for conversation sentiment analysis. The use of graph convolutional networks to capture the inter-speaker dynamics in a dialogue was explored by [28]. They utilized the dependencies among speakers to capture the contextual dynamics in an efficient way. In another work, AGHMN [29] proposed a hierarchical memory network with an attention mechanism to capture the essence of the dialogue in order to get a better understanding of the emotional dynamics of the speaker. TL-ERC [30] exploited the learned parameters

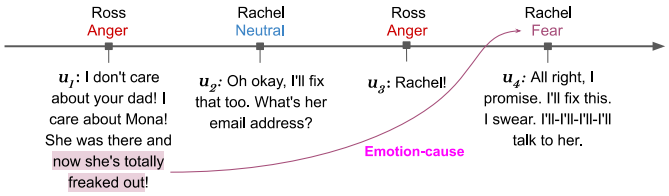
of a dialogue generation module for emotion recognition through the transfer learning setup. Recently, DialogXL [31] adopted XLNet [38] model for ERC. They encoded the dialogue utterances and made use of dialogue-aware self-attention to exploit the dialogue semantics. A hierarchical gated recurrent unit framework involving two GRUs at different levels was employed by [33] where a lower-level GRU modeled the word-level inputs while an upper-level GRU captured the context at the utterance level. Further, [39] proposed a correction model for previous approaches called "Dialogical Emotion Correction Network (DECN)". The aim of this work was to improve upon the emotion recognition performance by automatically identify errors made by emotion recognition strategies. The authors proposed the use of a graphical network to model human interactions in dialogues. Another study [40] used graph to solve the problem of ERC. They proposed a conversational affective analysis model which combined dependent syntactic analysis and graph convolutional neural networks. A self-attention mechanism captures the most effective words in the conversation, followed by graph construction. The authors shows experiments on various datasets the report higher accuracy than previous methods.

Beyond emotion recognition. Most of the existing ERC systems do not account for the explainability of emotions. In an attempt to do so, [41] proposed the task of emotion-cause analysis. The task deals with identifying a span in the text responsible for a specific emotion. For instance, we observe two emotions in the sentence '*The queue was so long, but at last I got vaccinated*' – *joy* and *disgust*. The task aims at identifying the phrases '*the queue was long*' for *disgust* and '*I got vaccinated*' for *joy*. Following this work, [42] investigated the use of linguistic phenomenon by proposing an SVM-based model for emotion-cause identification. Xia et al. [43] proposed another task- emotion-cause pair extraction. This task tried to extract the potential pairs of emotions and the corresponding causes in a document. The proposed a two-step approach where, first, individual emotion extraction and cause extraction are performed via multi-task learning and then emotion-cause pairing and filtering are done. In another one of their work [44] proposed a joint emotion-cause extraction framework which consisted of two encoders. A RNN based encoder was employed to get the word-level representations while a Transformer based encoder was applied to to learn the correlation between multiple clauses in a document. They also encoded relative position and global predication information that they claim helped capture the causality between clauses. Recently, [45] extended the task of emotion-cause extraction for conversation and released a dataset called RECCON containing $1,000+$ dialogs accompanied by $10,000$ emotion-cause pairs.

How is our task different? EFR represents a novel paradigm in NLP as it deals explicitly and quantitatively with identifying emotion instigators. While word embeddings may contain some implicit information about utterance meaning and emotion dynamics, they provide no explainability for an emotion-flip, and hence, cannot be used as a potential feedback mechanism to a response generator. Additionally, the two tasks, namely emotion-cause extraction in conversation and emotion-flip reasoning (EFR), may seem similar at an abstract level; nonetheless, they differ considerably at the surface level. While emotion-cause extraction in conversation aims to extract a text span that acts as grounds for the elicited emotion, EFR is a more speaker-specific task that highlights the instigators responsible for a "flip" in the speakers' emotion. In our case, the instigators (or causes) for an emotion flip come from a finite set of predefined labels, in contrast with the infinite possibilities of a span that the emotion-cause extraction task can extract. In order to reinforce the difference between the two tasks, we show a sample dialogue in Figure 3 from MELD-I with annotated EFR and emotion-cause labels. It can be observed that the reason behind the emotion *fear* in utterance u_4 comes from utterances u_1 and u_3. On the other hand, the emotion flip from *neutral* to *fear* (from utterance u_2 to u_4) was triggered by the utterance u_3 because of the feelings of *annoyance* and *scold* being instigated in the target speaker.



(a) Emotion flip reasoning



(b) Emotion-cause extraction

Fig. 3: A sample dialogue to illustrate the difference between emotion-cause extraction in conversation and emotion-flip reasoning.

III. Dataset

In this work, we use the existing MELD dataset [46] and augment it with EFR labels to get the new **MELD-I** (abbreviation of MELD-Instigators) dataset. MELD is a chit-chat dialogue dataset compiled from a famous TV series, F.R.I.E.N.D.S.[1] It contains $1,433$ multi-party dialogues with $13,708$ utterances. Each utterance is associated with one of the six Ekman's [13] emotion labels – *anger*, *fear*, *disgust*, *sadness*, *joy*, and *surprise*. Additionally, the absence of emotion is marked with the *neutral* label. We include new ground-truth labels for EFR to get MELD-I. We call the proposed MELD-I as a "new" dataset primarily due to the ***manually annotated*** instigator labels. Specifically, we propose new labels and convert the original MELD dataset into instances based on emotion-flips as illustrated in Table III resulting in 1161 dialogue instances. We manually annotate all these instances with the proposed instigator labels. However, since we use the

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

dialogues and emotion labels from the MELD dataset, we keep the name of the new dataset derived from it- MELD-I.

A. Instigator Labels

To understand the emotional dynamics of the speakers in a conversation, it is imperative to reason out any change/flip of the emotion of any speaker. Following the Cognitive Appraisal Theory by Lazarus et al. [9], which states that emotions are a result of appraisals, we aim to identify these appraisals for each emotion flip in the dialogue. These instigators follow the following properties:

- The instigators need not be unique to an emotion flip. For example, *threat* can instigate the emotion flip *joy* $\rightarrow$ *fear* as well as the flip *anger* $\rightarrow$ *fear*.
- An emotion flip need not necessarily arise from the same set of instigators. For example, the emotion flip *neutral* $\rightarrow$ *fear* can be caused by *threat* and *challenge* in different situations.
- There can be more than one instigator corresponding to a single emotion flip. For example, for the emotion flip *neutral* $\rightarrow$ *fear*, the instigator can be both *threat* and *challenge*.
- The instigators cannot be emotions themselves. For example, for the emotion flip *neutral* $\rightarrow$ *surprise*, the instigator cannot be *joy*.

We organize these instigators in a 2-level hierarchy. The first level presents a coarser representation of the instigators with 14 labels, while the second level defines all 27 instigators as fine-grained representation. We present the hierarchy of instigators and their definitions in the supplementary. Further, these instigators can be divided into three sets, based on the target emotion they can instigate- positive, negative, and neutral. A detailed discussion can be found in the supplementary.

B. Annotation Process

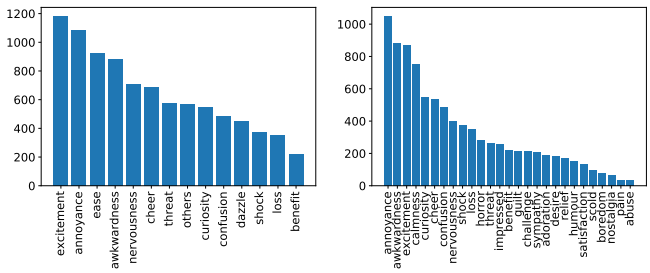
The first step in our annotation process is the instance creation for each emotion flip, followed by the trigger identification and instigator labeling. Table I presents the outcome of the annotation process for the example shown in Figure 1. We explain these steps in detail below.

1) **Instance creation:** For each emotion flip of a speaker, we create a new instance. The instance contains the utterances from the beginning of the dialogue till the target utterance (emotion flipped utterance). Utterances u_3 and u_5 are the target utterances in Table I, and utterances $\langle u_1, u_2, u_3 \rangle$ and $\langle u_1, u_2, u_3, u_4, u_5 \rangle$ are the respective candidate triggers for the target utterances. Among these candidates, u_2, u_3 are the triggers for the target u_3, while utterance u_4 instigates the emotion flip in the target u_5. Intuitively, the last utterance of each instance is the target utterance – the location of emotion flip. Note that we remove all such dialogues from MELD that do not contain an emotion flip which removed 271 dialogues from the set.
2) **Trigger identification:** After creating an instance for each emotion flip, we identify a set of trigger utterances that cause the emotion to flip at the target. We mark each utterance that acts as a trigger as 'Yes' and the ones not contributing as 'No'. The two instances in Table I have utterances $\langle u_2, u_3 \rangle$ and $\langle u_4 \rangle$ as triggers for the target utterances u_3 and u_5, respectively.
3) **Instigator labeling:** Finally, we assign one or more instigator labels to each trigger utterance corresponding to the target utterance. For example, as presented in Table I, we assign '*nervousness*' and '*adoration*' instigators to the trigger utterances u_2 and u_3, respectively, for the target u_3. Similarly, for the target utterance u_5 in Table I, we annotate the trigger u_4 with two instigator labels '*annoyance*' and '*challenge*'. It is evident that the instigator identification is a multi-label problem.

We employ the services of three annotators[2] to annotate MELD-I – two of them in the first stage of annotation, while the service of the third expert was sought to resolve any disagreement. We compute Krippendorff's Alpha inter-annotator agreement [47] to measure the consistency in the annotation. For trigger identification, we obtain the inter-annotator agreement between annotators A and B as $\alpha_{AB} = 0.817$, between annotators B and C as $\alpha_{BC} = 0.820$, and between annotators A and C as $\alpha_{AC} = 0.811$. We take the average of these three to get the overall agreement rating, i.e., $\alpha = 0.816$. For the instigator annotation, $\alpha_{AB} = 0.511$, $\alpha_{BC} = 0.545$, and $\alpha_{AC} = 0.540$. The average agreement comes out to be $\alpha = 0.532$. The low value for the latter is attributed to the multi-label characteristic of the task.

C. Dataset Statistics

We show a brief statistic of MELD-I in Table II along with the distribution of emotion flips. We also show the distribution of instigators in Figure 4. We can observe from Figure 4b that the distribution of fine-grained instigator labels is skewed towards a few of the instigators. As an attempt to reduce the skewness, we group similar instigator labels and obtain a reduced set of 14 instigators in the coarse-grained setup (c.f. Figure 4a).



(a) Coarse-grained (b) Fine-grained

Fig. 4: Distribution of EFR instigators in MELD-I.

It is interesting to note that emotion flips can be divided into two categories – **positive** emotion flips ({*anger*, *fear*, *disgust*, *sadness*} $\rightarrow$ {*joy*, *surprise*, *neutral*}) and **negative** emotion flips ({*joy*, *surprise*} $\rightarrow$ {*anger*, *fear*, *disgust*, *sadness*, *neutral*}). The flips {*neutral*} $\rightarrow$ {*joy*, *surprise*} are also considered

[2]They are NLP researchers or linguistics by profession; their age ranges between $30 - 45$ years.

TABLE I: Dataset development for an instance shown in Figure 1. (a) Original dialogue from MELD; (b,d) Two instances corresponding to the two emotion flips in (a); (c,e) Trigger and instigator annotations for both instances.

(a) An example dialogue from MELD.

	Speaker	Utterance	Emotion	
u_1	Ross	No! No sir umm, she means a lot to me. I mean, I care I-I love Rachel.	Fear	
u_2	Mona	What?!	Surprise	
u_3	Ross	Oh but not that way. I mean I mean I'm not in love with her. I love her like a, like a friend.	Joy	Emotion flip - 1
u_4	Dr. Green	Oh really? That's how you treat a friend? You get her in trouble and then refuse to marry her?	Anger	
u_5	Ross	Hey! I offered to marry her!	Anger	Emotion flip - 2

(b) Instance - 1.

	Speaker	Utterance	Emotion
u_1	Ross	No! No sir umm, she means a lot to me. I mean, I care I-I love Rachel.	Fear
u_2	Mona	What?!	Surprise
u_3	Ross	Oh but not that way. I mean I mean I'm not in love with her. I love her like a, like a friend.	Joy

(c) MELD-I Annotation: Trigger/Instigator.

	Speaker	Utterance	Emotion	Trigger	Instigator
u_1	Ross	No! No sir umm, she means a lot to me. I mean, I care I-I love Rachel.	Fear	No	-
u_2	Mona	What?!	Surprise	Yes	Nervousness
u_3	Ross	Oh but not that way. I mean I mean I'm not in love with her. I love her like a, like a friend.	Joy	Yes	Adoration

(d) Instance - 2.

	Speaker	Utterance	Emotion
u_1	Ross	No! No sir umm, she means a lot to me. I mean, I care I-I love Rachel.	Fear
u_2	Mona	What?!	Surprise
u_3	Ross	Oh but not that way. I mean I mean I'm not in love with her. I love her like a, like a friend.	Joy
u_4	Dr. Green	Oh really? That's how you treat a friend? You get her in trouble and then refuse to marry her?	Anger
u_5	Ross	Hey! I offered to marry her!	Anger

(e) MELD-I Annotation: Trigger/Instigator.

	Speaker	Utterance	Emotion	Trigger	Instigator
u_1	Ross	No! No sir umm, she means a lot to me. I mean, I care I-I love Rachel.	Fear	No	-
u_2	Mona	What?!	Surprise	No	-
u_3	Ross	Oh but not that way. I mean I mean I'm not in love with her. I love her like a, like a friend.	Joy	No	-
u_4	Dr. Green	Oh really? That's how you treat a friend? You get her in trouble and then refuse to marry her?	Anger	Yes	Annoyance, Challenge
u_5	Ross	Hey! I offered to marry her!	Anger	No	-

as positive emotion flips whereas {*neutral*} $\rightarrow$ {*anger*, *fear*, *disgust*, *sadness*} are considered as negative emotion flips. Considering the above categorization of emotion flips, we observe that not all instigators can result in all emotions flips. For example, it is improbable for a person to feel *joy* because of *guilt* – for an emotion flip with the target emotion *joy*, the instigator can almost never be *guilt*. Our observation of the annotated dataset is in line with this phenomenon.

Consequently, we divide our instigator labels into three sets – positive, negative, and ambiguous (c.f. Section V-E). We observe that for a positive emotion flip, only the instigators belonging to the positive and ambiguous set of instigators are responsible. Similarly, for a negative emotion flip, the negative and ambiguous sets are applicable.

TABLE II: Statistics of the dataset, MELD-I.

(a) MELD-I dataset for Emotion Flip Reasoning (EFR)

Split	#Dialogue with Flip	#Utterance with Flip	#Triggers
Train	834	4001	5262
Dev	95	427	495
Test	232	1002	1152

(b) Frequency of emotion flips with respect to emotions

Source \ Target	Disgust	Joy	Surprise	Anger	Fear	Neutral	Sadness
Disgust	0	24	30	47	6	76	13
Joy	34	0	169	86	42	665	81
Surprise	39	186	0	137	32	400	70
Anger	37	96	104	0	20	318	99
Fear	7	20	23	45	0	87	27
Neutral	84	616	487	370	103	0	257
Sadness	17	78	60	72	28	238	0

IV. METHODOLOGY

This section dwells on our proposed model, TGIF, to identify the EFR instigator labels for each emotion flip. The instigator identification task is a multi-label instance classification problem, as more than one instigator is possible for each trigger. TGIF models the global utterance sequence (*aka.* dialogue context) and speaker dynamics to capture the underlying semantics in the dialogue. Moreover, considering the strong relationship of emotion with our task, we also encode the emotion sequence of the utterances in TGIF. In total, TGIF has four submodules that exploit the global and speaker-specific dialogue and emotion dynamics – Global Utterance Sequence (GUS), Global Speaker Sequence (GSS), Global Emotion Sequence (GES), and Speaker-Specific Emotion Sequence (SSES). Finally, we combine the outputs of these four modules through a series of fully-connected layers followed by a 14/27 neurons sigmoid layer for both coarse-grained and fine-grained instigator identifications. Furthermore, at the penultimate layer, we apply an output mask to filter out the improbable instigator labels for the underlying emotion flip. The output mask assists the model in focusing on the probable labels and blocks the gradients for the unlikely labels to propagate back to the network. Below, we describe each module of TGIF in detail. Figure 5 presents the architecture of TGIF.

a) **Global Utterance Sequence (GUS):** The principle information about a dialogue lies in the utterances spoken in it. Thus, we employ GUS to encode the utterance sequence. We use a Transformer [48] encoder to extract a hidden representation h_i^u for each utterance u_i. For each, utterance, $u_i^{s_i}$, in the dialogue, we get an encoded vector, h_i^u, after this state, i.e. $\forall u_i^{s_i}, h_i^u = T_u(u_i^{s_i})$. Thus, h_i^u represents the context aware representation of the i^{th} utterance of the dialogue D.

b) **Global Emotion Sequence (GES):** In this module, we employ a single-layer GRU [49], $gGRU$, to capture the global emotion sequence of the dialogue. We hypothesize

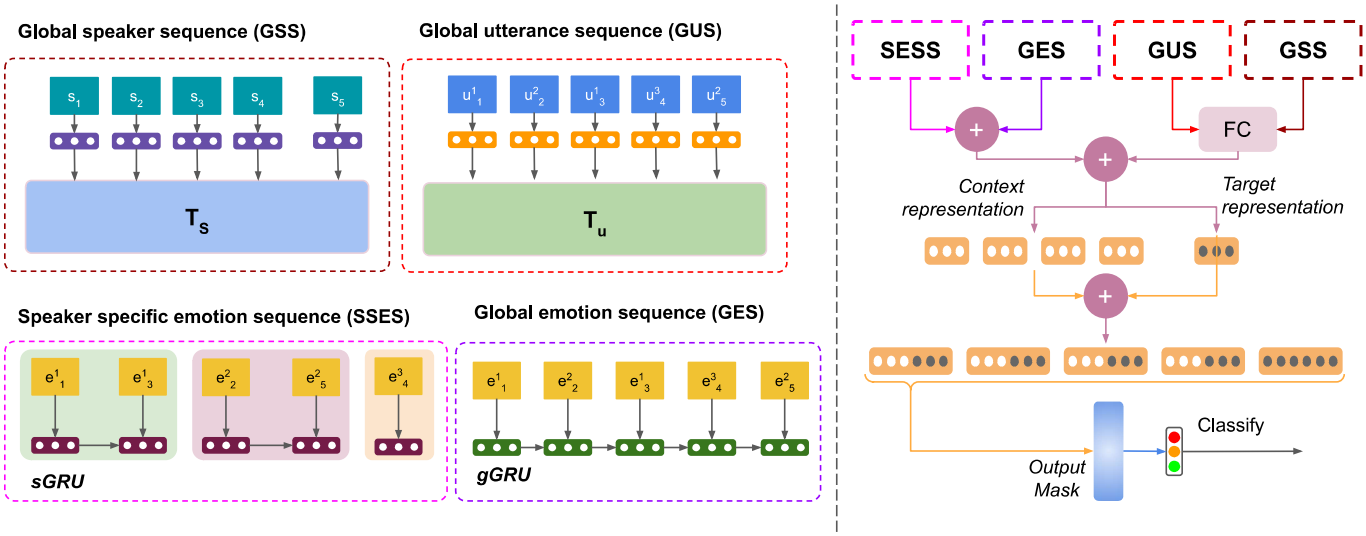


Fig. 5: The proposed TGIF architecture. Input: $\{\langle u_1, s_1, e_1\rangle, \langle u_2, s_2, e_2\rangle, \langle u_3, s_1, e_3\rangle, \langle u_4, s_3, e_4\rangle, \langle u_5, s_2, e_5\rangle\}$, where $\langle u_i, s_j, e_i\rangle$ represents the utterance u_i by a speaker s_j and its associated emotion e_i. Target (emotion flipped) utterance: $\langle u_5, s_2, e_5\rangle$ as $e_2 \neq e_5$.

that the knowledge of emotion sequence would assist the model in capturing a high-level snapshot of the emotion flow among speakers. We feed the emotion sequence of the dialogue, $\{e_1, e_2, ..., e_t\}$, as input to the GRU where each emotion e_i is represented by a one-hot vector of dimension 7. As a result, we obtain the hidden representation as follows: $[h^e_1, .., h^e_2, .., h^e_t] = gGRU(e_1, e_2, .., e_t)$.

c) **Speaker-Specific Emotion Sequence (SSES):** Each emotion flip is associated with a speaker. Thus, we hypothesize that the sequence of emotions at the speaker level is crucial and would exploit the emotion dynamics of each speaker considering the target speaker. Moreover, it would distinguish between the emotional states of the target speaker and other speakers. To achieve this, we employ separate GRUs for each speaker in an instance.

For example, if there are three distinct speakers in an instance (c.f. instance 2 in Table I), we learn three GRUs. For each speaker, we extract its emotion from the dialogue and create the input for GRUs as follows. Let an instance with five utterances of three distinct speakers be given as $\{\langle u_1, s_1, e_1\rangle, \langle u_2, s_2, e_2\rangle, \langle u_3, s_1, e_3\rangle, \langle u_4, s_3, e_4\rangle, \langle u_5, s_2, e_5\rangle\}$, where u_i and e_i denote the utterance and associated emotion at turn i by speaker s_j. We compile three inputs for each speaker as $\{e_1, e_3\}$, $\{e_2, e_5\}$, and $\{e_4\}$, and feed them to three speaker-specific GRUs (sGRU).

$$[\hat{h}^e_1, \hat{h}^e_3] = sGRU_1(e_1, e_3)$$
$$[\hat{h}^e_2, \hat{h}^e_5] = sGRU_2(e_2, e_5)$$
$$[\hat{h}^e_4] = sGRU_3(e_4)$$

Finally, we combine the hidden representations of GRUs by arranging them in the dialogue order for further processing, i.e., $\hat{H} = [\hat{h}^e_1, \hat{h}^e_2, \hat{h}^e_3, \hat{h}^e_4, \hat{h}^e_5]$.

d) **Global Speaker Sequence (GSS):** To explicitly capture the speaker information, their reactions with respect to other speakers, and their relationships, we also propose to encode the speaker sequence in TGIF. To capture the different reactions of a speaker with respect to the utterance of other speakers, we capture the speaker sequence by employing another Transformer encoder, T_s, which takes as input the sequence of speakers where each speaker is represented by a one-hot encoded vector. Each speaker goes through the Transformer encoder to give a speaker sequence aware representation, h^s_i, i.e. $\forall s_i, h^s_i = T_s(s_i)$. After this, we have a speaker sequence aware representation for each speaker, h^s_i, in the dialogue.

e) **Fusion:** We fuse the outputs of the above four submodules in two steps. In the first step, we combine the dialogue-level utterance and speaker sequence to obtain a global view of the conversation through a fully-connected layer. In parallel, we combine the dialogue and speaker-level emotion dynamics to get the essence of the flow of emotions in the conversation. Subsequently, in the second step, we concatenate the two representations. As the effect of an utterance on the final emotion changes with the change of the target utterance, we append the target representation to each utterance before feeding it to the output layer for prediction. We can justify the appending operation through the example shown in Table I. It can be observed that utterance u_2 is present in both instances; however, it is the trigger only in the first instance. Moreover, in Figure 2, the same trigger utterance u_3 resulted in an emotion flip from *neutral* to *sadness* in the first instance, while it causes the speaker's emotion change to *surprise* from *joy* for the second instance. Finally, we apply gradient masking for the improbable instigators.

V. Experiments and Results

We perform experiments for both the granular levels – coarse-grained and fine-grained. In the fine-grained setup, we observe a few instigator labels with a very low count. Since these labels are few in number, the model does not have sufficient evidence to learn a mapping from the input to such labels. Consequently, we compile another coarse-grained setup where we merge all instigator labels with count < 250 into a set, called '*other*'. As a result, in total, we have three setups – one fine-grained

with 27 instigator labels) and two coarse-grained (definition-based (c.f. Section III) and count-based) with 14 instigator labels each. In all three setups, we employ sigmoid neurons with focal loss [50] for multi-label classification. We select the traditional precision, recall, and F1-score as our metrics of choice thus ensuring that our evaluation is consistent with existing practices and establishes a universal benchmark.

A. Development Phase

To find the best configuration for TGIF, we investigate the effect of each module in the development phase. We start with the GUS module as the backbone network and subsequently introduce other modules (GES, GSS, and SESS) in an incremental fashion. Table III illustrates the results we obtain. Looking at the fine-grained setup, we notice a performance increase of 1.6% in weighted F1 when we add the GES module to the backbone model. This performance increase is coherent with our argument that the inclusion of emotional information will help the model in learning a better mapping function. Additionally, the incorporation of speaker-specific modules (GSS and SESS) gives a performance boost of 0.6% and 1.6%, respectively. We use the model consisting of all four submodules as our final architecture since it yields the best performance on the development set. After fine-tuning the hyperparameters during the development phase, we fix the configuration and evaluate TGIF on the test set.

TABLE III: Results (W-F1) of fine-tuning on the development set. It shows the effect of each module when incorporated in TGIF. We obtain the best results when all four submodules are employed (last row).

Model	Coarse-grained		Fine-grained
	Defn-based	Count-based	
GUS	41.4	37.0	29.9
+ GES	41.9	37.1	31.5
+ GSS	42.1	37.8	32.1
+ SESS (TGIF)	**42.7**	**38.5**	**33.1**

B. Baselines and Comparative Study

Since the problem of instigator classification for EFR is novel, we adapt various related existing systems for comparison. Note that all these systems are recent state-of-the-arts and designed especially for the task of emotion recognition in conversation (ERC).

1) **DialogueGCN** [28]: It exploits self and inter-speaker dependency to recognize emotion in conversations. It captures utterance embeddings with GRUs, followed by a graph convolution network (GCN) to leverage the speaker-level context, which is finally used for the task of emotion classification.
2) **AGHMN** [29]: It uses an attention-based GRU to monitor the flow of information through a hierarchical memory network. The attention weights are calculated over the contextual utterances in the conversation and combined for the final classification.
3) **TL-ERC** [30]: It is a transfer learning-based framework for emotion recognition in conversation. The weights of a hierarchical encoder-decoder model, trained for the dialogue generation task, are used for emotion classification.
4) **DialogXL** [31]: It modifies XLNet and evaluates it for emotion recognition in conversation (ERC). The authors changed the segment-level recurrence mechanism to an utterance-level recurrence mechanism so that XLNet could be mapped to a dialogue setting and incorporated dialogue-aware self-attention to capture the intra- and inter-speaker dependencies in a conversation.
5) **BERT** [51]: It is basically an encoder stack of transformer architecture [52].

TABLE IV: Comparative results on coarse-grained and fine-grained instigators. All the metrics are weighted average over all instigator classes.

Model	Coarse-grained						Fine-grained		
	Defn-based			Count-based					
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
AGHMN	7.6	20.4	11.07	8.5	25.4	12.73	15.1	17.6	16.3
TL-ERC	9.6	49.0	16.6	14.4	54.8	21.7	7.1	33.0	12.8
DGCN	12.5	67.0	19.8	18.5	70.2	27.5	10.5	67.2	17.5
DialogXL	7.3	37.5	12.22	8.8	43.7	14.64	9.8	34.2	15.3
BERT	18.3	62.9	27.2	17.5	59.2	26.3	14.8	55.1	21.7
TGIF	**24.3**	58.6	**31.6**	**28.3**	63.4	**37.5**	**26.5**	55.6	**33.3**

Similar to TGIF, we perform instance-wise experiments with output masking for each baseline. That is, all the improbable instigators are masked. Moreover, since we provide emotion labels as input to our model, we do the same with the baselines.

Table IV shows that TGIF outperforms all baselines with reasonable margin across all setups. In the definition-based coarse-grained setup, we obtain 11.07%, 16.6%, 19.8%, 12.2%, and 27.2% weighted-F1 for AGHMN, TL-ERC, DGCN, DialogXl, and BERT respectively. In comparison, TGIF yields 31.6% W-F1 in the same setup – an increment of 4.4 points over the best performing baseline (BERT). We observe a similar trend for the count-based coarse-grained setup with TGIF and the best baseline (DGCN) reporting 37.5% and 27.5% W-F1, respectively – a difference of 10 points. In the fine-grained setup, the performance of the baselines (ranging between $12.8\% - 21.7\%$) are significantly inferior to TGIF (33.3%). It suggest that TGIF also accounts for the increase in instigator labels more efficiently than the existing baselines. TGIF beats all considered baselines in every setting but at the same time reports a weighted F1 score on the lower side indicating the difficulty of the problem statement.

C. Result Analysis

As we can observe from the distribution of EFR instigators (c.f. Figure 4), there is a significant label skewness. To inspect the learning of TGIF for individual labels, we analyze the results of top-3 (majority) and bottom-3 (minority) instigator labels w.r.t. the number of training instances in MELD-I. The top-3 labels are *nervousness*, *awkwardness*, and *excitement* in the coarse-grained setup, and *annoyance*, *awkwardness*, and *excitement* in the fine-grained setup. Similarly, The bottom-3 labels in the coarse-grained setup are *shock*, *dazzle*, and *threat*, while instigators *nostalgia*, *pain*, and *boredom* are the three least occurring labels in the fine-grained setup.

TABLE V: Fine-grained analysis: Actual and predicted instigator labels for an EFR instance. TGIF predicts one and two correct instigator labels for the two trigger utterances, u_4 and u_5, respectively. In each case, it wrongly predicts one instigator. In contrast, BERT reports a high percentage of *false positives*.

	Speaker	Utterance	Emotion	Trigger	Instigator		
					Gold	Prediction	
						TGIF	BERT
u_1	Monica	Yeah, but without the costumes.	neutral	No	-	-	-
u_2	Phoebe	Oh.	neutral	No	-	-	-
u_3	Joey	And it's not fake, it's totally brutal.	neutral	No	-	-	-
u_4	Chandler	Yeah, it's two guys in a ring, and the rules are: They are no rules.	neutral	Yes	confusion, curiosity	confusion, shock	excitement, nervousness, shock
u_5	Monica	So you can like, bite, and pull people's hair and stuff?	surprise	Yes	confusion, shock	confusion, shock, curiosity	curiosity, shock

TABLE VI: Class-wise comparative analysis (F1-score) for the top-3 (majority) and bottom-3 (minority) classes. ⟨*Ner: Nervousness*, *Awk: Awkwardness*, *Exc: Excitement*, *Ann: Annoyance*, *Shk: Shock*, *Daz: Dazzle*, *Tht: Threat*, *Nos: Nostalgia*, *Bor: Boredom*⟩.

Model	Top-3 Majority						Bottom-3 Minority					
	Coarse-grained			Fine-grained			Coarse-grained			Fine-grained		
	Ner	Awk	Exc	Ann	Awk	Exc	Shk	Daz	Tht	Nos	Pain	Bor
AGHMN	11.2	10.1	8.6	17.8	12.8	15.3	3.9	4.2	3.2	**13.4**	7.2	**16.9**
TL-ERC	23.0	23.3	18.8	14.9	23.5	14.6	2.2	10.6	5.4	6.6	0.6	1.8
DGCN	28.9	28.4	28.3	24.8	26.9	23.2	10.5	9.8	6.0	0.0	**7.6**	2.1
DialogXL	12.2	14.7	12.6	13.9	11.9	11.1	2.1	3.6	4.2	6.4	3.2	2.9
BERT	36.0	26.8	**35.3**	38.9	26.8	36.1	10.7	2.5	0.0	3.2	2.0	0.0
TGIF	**37.8**	**35.7**	28.4	**53.5**	**35.8**	**56.7**	**35.1**	**18.6**	**9.8**	12.5	5.5	0.0

Tables VI report the results of TGIF and baselines for the majority and minority classes, respectively. As expected, the performance of each model for the majority classes is comparatively on the higher side of the spectrum than the performance on minority classes. Except for the instigators, *nostalgia*, *pain*, *boredom* in the fine-grained minority cases and *excitement* in the coarse-grained majority case, TGIF reports the best weighted-F1 for each case. The observed behaviour can be attributed to the fact that Transformer based architectures are data-hungry models, and thus they learn a better mapping for majority classes.

D. Qualitative Error Analysis

In order to perform qualitative error analysis, we take a sample dialogue from our test set and show the gold and predicted labels in Table V for the fine-grained setup. For the target utterance u_5, TGIF predicts *confusion* and *shock* instigators against the gold labels *confusion* and *curiosity* instigated by the trigger utterance u_4. Similarly, for the the trigger u_5, TGIF identifies two correct (*confusion* and *shock*) and one incorrect label (*curiosity*). An abstract view of the prediction suffices that the set of instigator labels for the emotion flip target u_5 (without regarding the triggers separately) is same as the set of gold labels. On the other hand, BERT (best baseline) commits many mistakes in both cases. It predicts one correct label for the trigger u_5 but no correct instigator for the trigger u_4. It can be observed that BERT gives precision scores of 0% for the first trigger while a precision of 50% is observed for the second trigger. Recall value also comes out to be 0% and 50% for the two triggers, respectively. In comparison, TGIF obtains moderate scores in both cases, i.e., recall $= 50.0\%$; precision $= 50.0\%$ in the first case and recall $= 100.0\%$; precision$= 66.7\%$ in the second case. A similar trend is observed for coarse-grained instigator labels.

TABLE VII: Result analysis on directionality of flips for coarse-grained and fine-grained instigators. All the metrics are weighted average over all instigator classes.

Type of Flip	Coarse-grained						Fine-grained		
	Defn-based			Count-based					
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Negative to Positive	**27.3**	54.6	33.9	26.1	52.5	32.7	19.3	52.2	26.2
Positive to Negative	26.4	**61.2**	**35.1**	**28.0**	**59.1**	**35.4**	**32.2**	**58.1**	**38.8**

We note that, for a given target emotion, the BERT method consistently identifies the same instigators, giving little heed to the conversation context. In contrast, our approach takes both the target emotion and conversation context into account when identifying instigators, resulting in more accurate and nuanced predictions. Further, we show the zero-shot results of the proposed method in the supplementary.

E. Directionality of Triggers

In this work, we consider Ekman's emotion labels along with a label for no emotion (neutral). That is, we have six emotion labels, namely *disgust*, *joy*, *surprise*, *anger*, *fear*, and *sadness*. An emotion flip for a target speaker can occur between any two pair of emotions. In other words, we can have 42 possible emotion flips in a dialogue. Based on the source-target emotion pairs, we analyse the effect of directionality of emotion flips in MELD-I. We show the frequency of emotion flips with respect to the source-target emotion pairs in Table II. Cell (i, j) in the table represents the number of flips in MELD-I where the source emotion is e_i, and the resultant or target emotion is e_j. As discussed in Section III-C, there can be two types of emotion flips – *positive* and *negative*. Here, we see how the emotion flips frequency and instigators are dependant on the type of flips. Based on our ground-truth EFR labels, there are a total of 2612 positive emotion flips and 2818 negative emotion flips. Out of these flips, the flip *neutral* $\rightarrow$ *joy* is the most prominent positive emotion flip whereas the flip, *joy* $\rightarrow$ *neutral* is the most prominent negative emotion flip.

Apart from the emotion flips which have opposite polarities at both ends, we can also have intra-polarity flips. For instance, flips like *anger* $\rightarrow$ *fear* is a negative to negative emotion flip, while *joy* $\rightarrow$ *surprise* is a positive to positive emotion flip. We see, for the intra-polarity cases, that the flip *surprise* $\rightarrow$ *joy* and the flip *anger* $\rightarrow$ *sadness* are the most prominent intra-positive and intra-negative flips, respectively. We also observe that most of the flips that result in a negative emotion (*anger*, *disgust*, *fear*, *sadness*) originate from *joy*. On the other hand, the flips that result in a positive emotion (*joy*, *surprise*) originate from

neutral. We also observe that, for positive emotion flips, the top-3 frequent instigators are *excitement*, *cheer*, or *impressed*. For negative emotion flip, *awkwardness*, *loss*, or *annoyance* are the more frequent instigators.

In addition, we check the performance of our models on the two most prominent types of flip directions – positive to negative and negative to positive. We show these results in Table VII and observe that positive to negative flips are better predicted by our model for all the classes of instigators. This can be attributed to the fact that our data contains more negative emotions, thus containing more negative instigators. Consequently, our model is able to learn those instigators in a better fashion. This result is encouraging as it is an indication that with more data, our model will be able to learn the instigators in a better way.

F. Generalizability of TGIF

In order to emphasise the relevance of EFR and evaluate the generalizability of the proposed methodology, TGIF, we perform a zero-shot experiment. We consider IEMOCAP [53] which consists of emotion annotated conversations on 16 topics. We randomly sample 15 conversations to construct emotion flip instances, with triggers identified as shown in Table III of the main text. We then task TGIF and BERT, the best baseline, with predicting the emotion flip instigators. After collecting the predictions, we asked 20 human annotators to rate them based on correctness, completeness, and preference (TGIF vs. BERT). The cumulative results in Table VIII indicate that while TGIF outperforms BERT, the latter is comparable in terms of completeness.

TABLE VIII: Human Evaluation Results on IEMOCAP [53] in a zero-shot setting. Scores are average across all evaluators.

	Correctness	Completeness	Prefered Instigator set
BERT	2.67	3.42	25%
TGIF	**3.21**	**3.44**	**75%**

VI. Conclusion

This paper focused on explaining the reasons for a flip of a speaker's emotion in a conversation. Our interest lies in revealing which cognitive appraisal instigated the flip. We introduced Emotion Flip Reasoning (EFR), which aims to identify these appraisals (or instigators) responsible for an emotion flip. To address EFR, we prepared a new dataset, MELD-I, with the annotations of responsible utterances and respective appraisals. We further compiled the set of instigators in a two-level hierarchy – coarse-grained with 27 and fine-grained with 14 instigators. To benchmark the dataset, we proposed TGIF and performed extensive experiments for the instigator identification in both setups. The comparative analysis against five existing systems showed improvements in the range of $4-11\%$ W-F1 points. Moreover, TGIF adapted well in the fine-grained setup. The EFR task should motivate researchers to ponder more towards the explanability of the emotion dynamics involved in a conversation. The results of the EFR task can be exploited by dialogue agents in order to generate more empathetic response and it can also act as a reward function in the case of reinforcement learning. This is something we would like to explore in the future.

References

[1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[2] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from twitter text," *Journal of Computational Science*, vol. 36, p. 101003, 2019.

[3] L. Dini and A. Bittar, "Emotion analysis on twitter: The hidden challenge," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 3953–3958.

[4] M. S. Akhtar, A. Ekbal, and E. Cambria, "How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]," *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, pp. 64–75, 2020.

[5] N. Gupta, M. Gilbert, and G. D. Fabbrizio, "Emotion detection in email customer care," *Computational Intelligence*, vol. 29, no. 3, pp. 489–505, 2013.

[6] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2018. NIH Public Access, 2018, p. 2122.

[7] S. Kumar, A. Shrimal, M. S. Akhtar, and T. Chakraborty, "Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer," *Knowledge-Based Systems*, vol. 240, p. 108112, 2022.

[8] J. Mooren and I. Van Krogten, "Contributions to the history of psychology: Cxii. magda b. arnold revisited: 1991," *Psychological reports*, vol. 72, no. 1, pp. 67–84, 1993.

[9] R. S. Lazarus and S. Folkman, *Stress, appraisal, and coping*. Springer publishing company, 1984.

[10] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "Moel: Mixture of empathetic listeners," *arXiv preprint arXiv:1908.07687*, 2019.

[11] J. Shin, P. Xu, A. Madotto, and P. Fung, "Generating empathetic responses by looking ahead the user's sentiment," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7989–7993.

[12] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Information Fusion*, vol. 64, pp. 50–70, 2020.

[13] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[14] R. W. Picard, A. Wexelblat, and C. I. N. I. Clifford I. Nass, "Future interfaces: social and emotional," in *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, 2002, pp. 698–699.

[15] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[16] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, and C. Di Natale, "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure," *Knowledge-Based Systems*, vol. 63, pp. 68–81, 2014.

[17] L. Zhang, K. Mistry, S. C. Neoh, and C. P. Lim, "Intelligent facial emotion recognition using moth-firefly optimization," *Knowledge-Based Systems*, vol. 111, pp. 248–267, 2016.

[18] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, and X. Chen, "Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network," *Knowledge-Based Systems*, vol. 205, p. 106243, 2020.

[19] J. S. Y. Liew and H. R. Turtle, "Exploring fine-grained emotion detection in tweets," in *Proceedings of the NAACL Student Research Workshop*, 2016, pp. 73–80.

[20] O. Kaminska, C. Cornelis, and V. Hoste, "Nearest neighbour approaches for emotion detection in tweets," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021, pp. 203–212.

[21] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, pp. 245–263, 2022.

[22] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Emosec: Emotion recognition from scene context," *Neurocomputing*, pp. 174–187, 2022.

[23] Y. Li, T. Zhang, and C. L. P. Chen, "Enhanced broad siamese network for facial emotion recognition in human–robot interaction," *IEEE Transactions on Artificial Intelligence*, pp. 413–423, 2021.

[24] G. Assunção, B. Patrão, M. Castelo-Branco, and P. Menezes, "An overview of emotion in artificial intelligence," *IEEE Transactions on Artificial Intelligence*, pp. 867–886, 2022.

[25] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2594–2604.

[26] P. Zhong, D. Wang, and C. Miao, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 165–176. [Online]. Available: https://www.aclweb.org/anthology/D19-1016

[27] W. Li, W. Shao, S. Ji, and E. Cambria, "Bieru: bidirectional emotional recurrent unit for conversational sentiment analysis," *arXiv preprint arXiv:2006.00492*, 2020.

[28] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[29] W. Jiao, M. Lyu, and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8002–8009.

[30] D. Hazarika, S. Poria, R. Zimmermann, and R. Mihalcea, "Conversational transfer learning for emotion recognition," *Information Fusion*, vol. 65, pp. 1–12, 2021.

[31] W. Shen, J. Chen, X. Quan, and Z. Xie, "Dialogxl: All-in-one xl-net for multi-party conversation emotion recognition," *arXiv preprint arXiv:2012.08695*, 2020.

[32] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *ACL*, 2017, pp. 873–883.

[33] W. Jiao, H. Yang, I. King, and M. R. Lyu, "Higru: Hierarchical gated recurrent units for utterance-level emotion recognition," *arXiv preprint arXiv:1904.04446*, 2019.

[34] G. Tu, J. Wen, C. Liu, D. Jiang, and E. Cambria, "Context- and sentiment-aware networks for emotion recognition in conversation," *IEEE Transactions on Artificial Intelligence*, pp. 699–708, 2022.

[35] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 11 595–11 603.

[36] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, and Z. Yang, "A multi-view network for real-time emotion recognition in conversations," *Knowledge-Based Systems*, p. 107751, 2022.

[37] Z. Lian, B. Liu, and J. Tao, "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[38] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[39] Z. Lian, B. Liu, and J. Tao, "Decn: Dialogical emotion correction network for conversational emotion recognition," *Neurocomputing*, pp. 483–495, 2021.

[40] Y. Shou, T. Meng, W. Ai, S. Yang, and K. Li, "Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis," *Neurocomputing*, pp. 629–639, 2022.

[41] S. Y. M. Lee, Y. Chen, and C.-R. Huang, "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*, 2010, pp. 45–53.

[42] L. Gui, R. Xu, D. Wu, Q. Lu, and Y. Zhou, "Event-driven emotion cause extraction with corpus construction," in *Social Media Content Analysis: Natural Language Processing and Beyond*. World Scientific, 2018, pp. 145–160.

[43] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," 2019.

[44] R. Xia, M. Zhang, and Z. Ding, "Rthn: A rnn-transformer hierarchical network for emotion cause extraction," *arXiv preprint arXiv:1906.01236*, 2019.

[45] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, R. Ghosh, N. Chhaya, A. Gelbukh, and R. Mihalcea, "Recognizing emotion cause in conversations," *arXiv preprint arXiv:2012.11820*, 2020.

[46] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. [Online]. Available: https://www.aclweb.org/anthology/P19-1050

[47] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[49] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Oct 2017.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[53] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

Shivani Kumar is a PhD scholar at Indraprastha Institute of Information Technology Delhi (IIIT Delhi), India. She holds a Senior Research Fellowship and works in the domain of Natural Language Processing, primarily in the area of understanding and explaining various affects, like emotions, sarcasm, and humour, in conversational data.

Shubham Dudeja was a masters student at Indraprastha Institute of Information Technology Delhi (IIIT Delhi), India for the duration of this work. His research interest lies in Natural Language Processing and its sub-fields.

Md Shad Akhtar is currently an Assistant Professor at Indraprastha Institute of Information Technology Delhi (IIIT Delhi). His main area of research is NLP with a focus on the affective analysis. He completed his PhD form IIT Patna.

Tanmoy Chakraborty is an Associate Professor in the Dept. of Electrical Engineering, Indian Institute of Technology Delhi, India since September 2022. Before joining IIT Delhi, he served as an Associate Professor in the Dept of CSE, IIIT Delhi, India. He completed his postdoctoral research from University of Maryland, College Park after obtaining his PhD from the Dept. of CSE, IIT Kharagpur, India as a Google PhD scholar. His broad research interests include Natural Language Processing, Graph Neural Networks, and Social Computing. He is a senior IEEE member. More details about him can be found at tanmoychak.com.