Towards Emotion-aided Multi-modal Dialogue Act Classification

Tulika Saha[*], **Aditya Prakash Patra**[*], **Sriparna Saha, Pushpak Bhattacharyya**
Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
(sahatulika15,aditya.prakash.patra1997)@gmail.com
(sriparna.saha,pushpakbh)@gmail.com

Abstract

The task of Dialogue Act Classification (DAC) that purports to capture communicative intent has been studied extensively. But these studies limit themselves to text. Non-verbal features (change of tone, facial expressions etc.) can provide cues to identify DAs, thus stressing the benefit of incorporating multi-modal inputs in the task. Also, the emotional state of the speaker has a substantial effect on the choice of the dialogue act, since conversations are often influenced by emotions. Hence, the effect of emotion too on automatic identification of DAs needs to be studied. In this work, we address the role of *both* multi-modality and emotion recognition (ER) in DAC. DAC and ER help each other by way of multi-task learning. One of the major contributions of this work is a new dataset- multimodal Emotion aware Dialogue Act dataset called *EMOTyDA*, collected from open-sourced dialogue datasets. To demonstrate the utility of EMOTyDA, we build an attention based (self, inter-modal, inter-task) multi-modal, multi-task Deep Neural Network (DNN) for joint learning of DAs and emotions. We show empirically that multi-modality and multi-tasking achieve better performance of DAC compared to uni-modal and single task DAC variants.

1 Introduction

Dialogue Act Classification (DAC) is concerned with deciding the type i.e., communicative intention (question, statement, command etc.) of the speaker's utterance. DAC is very important in the context of discourse structure, which in turn supports intelligent dialogue systems, conversational speech transcription and so on. Considerable works have been done on classical Machine Learning (ML) based DAC (Jurafsky et al., 1997), (Stolcke et al., 2000), (Verbree et al., 2006), etc. and Deep Learning (DL) based DAC (Kalchbrenner and Blunsom, 2013), (Papalampidi et al., 2017), (Liu et al., 2017), (Ribeiro et al., 2019), (Ortega et al., 2019), (Saha et al., 2019) etc.

Humans are emotional entities. A speaker's emotional state considerably influences or affects its intended content or its pragmatic content (Barrett et al., 1993). An utterance such as "*Okay sure*" or "*Ya right*" (say) can be considered as "agreement" or- in case of sarcasm- "disagreement". For expressive DAs such as "greeting", "thanking", "apologizing" etc., the speaker's feeling or emotion can assist in recognizing true communicative intent and vice-versa. Thus, it is important to consider the speaker's emotion when deciding on the DA.

There is considerable work on ER (Cowie et al., 2001), (Jain et al., 2018), (Zhang et al., 2018), etc. and adapting the Virtual Agents (VAs) to act accordingly (Huang et al., 2018), (Zhou et al., 2018), (Fung et al., 2018), etc. But very little research has been done, that addresses the impact of emotion while deciding the DA of an utterance (Novielli and Strapparava, 2013), (Bosma and André, 2004). As DAs primarily dictate the flow of any dialogue conversation (be it human-human or human-computer), such synergy of ER and DAC is required. Research too has shown the benefit of utilizing the combination of text and nonverbal cues (Poria et al., 2017b), (Poria et al., 2017a) etc., for solving various Natural Language Processing (NLP) tasks. The main advantage of integrating other modalities to text is the usage of behavioral signs present in acoustic (vocal modulations) and visual (facial expression) modalities. In addition, the various modalities offer important signals to better identify the speaker's communicative intention and emotional state. This will in effect help create sturdy and more reliable DAC models.

In this paper, we study the influence of emotion on the identification of DAs, by utilizing the com-

[*]The authors have contributed equally.

bination of text, vocal modulations and facial expressions for task-independent conversations. DAC is our primary task, assisted by Emotion Recognition (ER) as an auxiliary task. We implement an attention based multi-modal, multi-tasking DNN to do joint modeling of DAC and ER. Also, we introduce a new dataset to help advance research in multi-modal DAC.

The key contributions of this paper are as follows: ***i.*** We curate a new dataset called EMOTyDA for facilitating multi-modal DAC research with high-quality annotations, including emotionally aided cues and conversational context features. We believe this dataset will advance research in multi-modal DAC; ***ii.*** We point to different scenarios where discrepancy in DAC is evident across different modalities, thus, showing the importance of multi-modal approaches to DAC; ***iii.*** We show using various instances, the usefulness of considering the emotional state of the user while identifying DAs. Consequently, we deduce that EMOTyDA will lead to a novel sub-task for future research: emotion aware DAC; ***iv.*** We propose an attention based (self, inter-modal, inter-task) multi-task, multi-modal DNN for jointly optimizing the DAC and ER task and show its benefit over single task DAC variants. Through this, we also establish that multi-modal DAC performs significantly better than uni-modal DAC.

2 Related Works

The tasks of ER and DAC are extensively explored.

Dialogue Act Frameworks: DAC has been investigated since late 90s (Reithinger and Klesen, 1997), (Stolcke et al., 1998) and early 2000's (Stolcke et al., 2000), (Grau et al., 2004). Much of this research, however, uses chat transcripts with only the text mode, due partly due to unavailability of multi-modal open-source dataset. In (Khanpour et al., 2016), authors apply stacked LSTM to classify speech acts. In (Kumar et al., 2018), the author developed a Hierarchical Network based approach using Bi-LSTMs and the CRF. A contextual self-attention system fused with hierarchical recurrent units was proposed by the authors of (Raheja and Tetreault, 2019) to develop a sequence label classifier. The authors of (Yu et al., 2019) proposed a method for the capture of long-range interactions that span a series of words using a Convolutional Network based approach. In (Saha et al., 2019), authors proposed several ML and DL based approaches such as Conditional Random Fields, clustering and word embeddings to identify DAs. However, all these works identify DAs by utilizing solely the textual modality without the use of emotional cues.

Emotion aware DAs. Within a multi-modal setting, little work is available in the literature to study the impact of emotional state in the evaluation of DAs. The effect of integrating facial features as a way of identifying emotion to classify DAs was examined by authors in (Boyer et al., 2011). They exhibited their work for tutorial dialogue session typically task-oriented and applied logistic regression to identify DAs. But they studied only the cognitive-affecting states such as *confusion* and *flow* as the emotional categories to learn DAs. In (Novielli and Strapparava, 2013), authors examined the impact of affect analysis in DA evaluation for an unsupervised DAC model. The authors made use of lexicon based features from *WordNet Affect* and *SentiWordNet* to map them with emotion labels to model the DAs in a LSA based approach. Authors of (Ihasz and Kryssanov, 2018), also inspected the impact of emotions mediated with intention or DAs for an *in-game* Japanese dialogue. Their goal was to construct DA-emotion combinations from the pre-annotated corpus. However, such stringent associations or dis-associations amongst DA-emotion pairs may not truly hold for real life conversations.

3 Dataset

To facilitate and enhance the research in multi-modal DAC assisted with user emotion, we introduce a new dataset (EMOTyDA) consisting of short videos of dialogue conversations manually annotated with its DA along with its pre-annotated emotions.

3.1 Data Collection

To gather potentially emotion rich conversations to explore its affect on DAC, we scanned the literature for existing multi-modal ER dataset. During our initial search, we obtained several multi-modal ER datasets which include *Youtube* (Morency et al., 2011), *MOUD* (Pérez-Rosas et al., 2013), *IEMOCAP* (Busso et al., 2008), *ICT-MMMO* (Wöllmer et al., 2013), *CMU-MOSI* (Zadeh et al., 2016), *CMU-MOSEI* (Zadeh et al., 2018) and *MELD* (Poria et al., 2019) etc. However, we zeroed down on IEMOCAP and MELD datasets for the further

investigations of our problem statement. The reason behind this choice was that remaining all the datasets mentioned above were particularly monologues involving opinions and product reviews. Whereas our research requires task-independent dyadic or multi-party conversations to analyze its full potential. Both these available datasets are not annotated for their corresponding DAs.

Also, benchmark DAC datasets such as Switchboard (SWBD) (Godfrey et al., 1992), ICSI Meeting Recorder (Shriberg et al., 2004) consist of text and audio-based conversations whereas TRAINS (Heeman and Allen, 1995) consist of solely text-based conversations with no emotional tags. HCRC Map Task corpus (Anderson et al., 1991) additionally encompasses audio modality with the transcripts but the corpus itself has task-oriented conversations and is not annotated for its emotion tags. It is to be noted that task-oriented conversations generally restrict the presence of diverse tags which are commonly encountered in task-independent conversations.

To the best of our knowledge, at the time of writing, we were unaware of any sizable and open-access DA and emotion annotated multi-modal dialogue data. Thus, MELD and IEMOCAP datasets have been manually annotated for the corresponding DAs to encourage and promote novel research on multi-modal DACs to build a multi-tasking system that allows DA and emotion for an utterance to be learned jointly.

3.2 Data Annotation

Over the years, SWBD-DAMSL tag-set comprising of 42 DAs developed by (Jurafsky, 1997) has been used widely for the task of DAC for task-independent dyadic conversation such as SWBD corpus. Thus, we use SWBD-DAMSL tag-set as the base for conceiving tag-set for the EMOTyDA dataset since both these datasets contain task-independent conversations. Of the 42 SWBD-DAMSL tags, 12 most commonly occurring tags have been used to annotate utterances of the EMOTyDA dataset. The choice of 12 tags is because of the limited length of the EMOTyDA dataset in comparison to the SWBD corpus. It stems from the fact that it is highly likely that many of the tags of the SWBD-DAMSL tag-set will never appear in the EMOTyDA dataset due to lesser number of utterances and lower diversity of occurrence of such fine-grained tags. The 12 most commonly occurring chosen tags are *Greeting* (g), *Question* (q), *Answer* (ans), *Statement-Opinion* (o), *Statement-Non-Opinion* (s), *Apology* (ap), *Command* (c), *Agreement* (ag), *Disagreement* (dag), *Acknowledge* (a), *Backchannel* (b) and *Others* (oth).

For the current work, we have selected a subset of 1039 dialogues from MELD amounting to 9989 utterances and the entire IEMOCAP dataset of 302 dialogues amounting to 9376 utterances to curate EMOTyDA dataset. Details of the original MELD and IEMOCAP datasets are provided in the Appendix 6.1. Three annotators who were graduate in English linguistics were accredited to annotate the utterances with the appropriate DAs out of the 12 chosen tags. They were asked to annotate these utterances by only viewing the video available considering the dialogue history without the information of the pre-annotated emotion tags. This was done so as to assure that the dataset does not get biased by specific DA-emotion pairs. The inter-annotator score over 80% was considered as reliable agreement. It was determined based on the count that for a given utterance more than two annotators agreed on a particular tag. To remove the discrepancy in the number of emotion tags for IEMOCAP and MELD datasets, we mapped the *joy* tag of the MELD to the *happy* tag of the IEMOCAP to finally settle on 10 tags from the IEMOCAP for the EMOTyDA dataset.

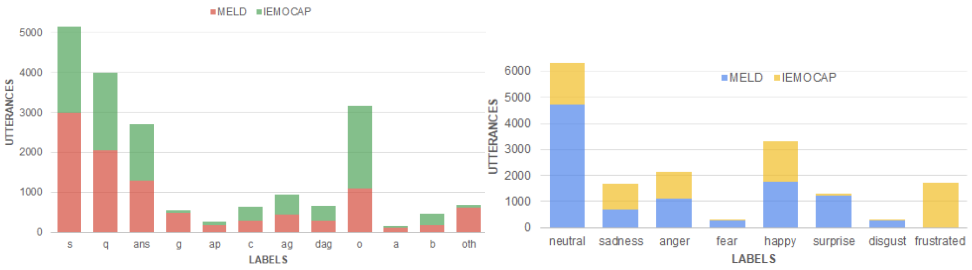
3.3 Emotion-DA Dataset: EMOTyDA

The EMOTyDA dataset[1] now comprises of 1341 dyadic and multi-party conversations resulting in a total of 19,365 utterances or annotated videos with the corresponding DA and emotion tags considering the dialogue history. The dataset contains approximately 22 hours of recordings. Source distribution and major speakers statistics of the dataset are shown in Figures 3a and 3b, respectively. Since DAC and ER tasks are known to exploit the contextual features, i.e., dialogue history (Yu et al., 2019) so, utterances in the dataset are accompanied with their corresponding contextual utterances, which are typically preceding dialogue turns by the speakers participating in the dialogue. Each of the utterances contains three modalities: video, audio, and text. All the utterances are even followed by their speaker identifiers. Table 1 shows few utterances along with the corresponding DAs and emotion la-

[1]The dataset with its DA and emotion tags will be made publicly available to the research community.

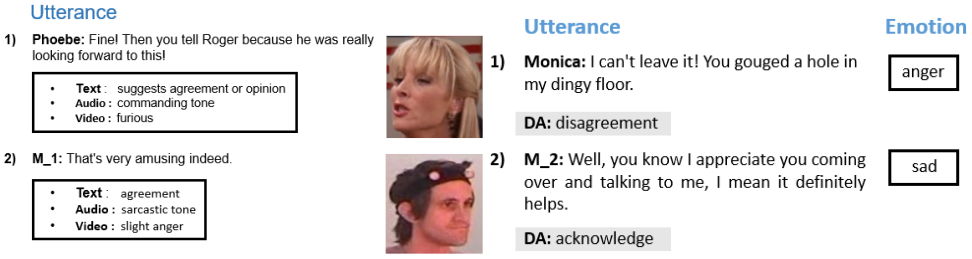
Speaker	Utterance	DA	Emotion
M_4	*Adders don't snap, they sting.*	dag	ang
Rachel	*Well, I just checked our messages and Joshua didn't call.*	s	sad
M_1	*That's very amusing indeed.*	dag	ang
Chandler	*Come on, pick up, pick up*	c	fear

Table 1: Example utterances from the EMOTyDA dataset with its corresponding DA and emotion categories



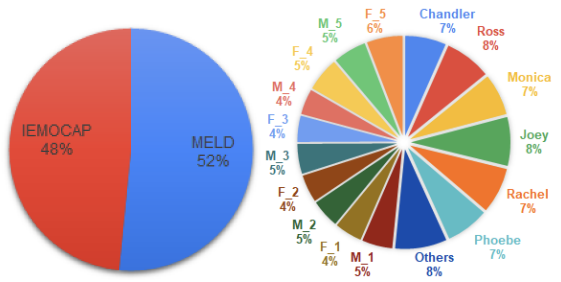
(a) (b)

Figure 1: Statistics across the datasets : (a) Distribution of DA labels, (b) Distribution of emotion labels.



(a) (b)

Figure 2: (a) Incongruent modalities in DAC, (b) Importance of emotion in DAC.



(a) (b)

Figure 3: Statistics : (a) Source across the dataset, (b) Overall speaker distribution.

bels from the proposed dataset. Distributions of DA and emotion labels across the source datasets are shown in Figure 1a and 1b, respectively.

3.4 Qualitative Aspects

In the current work, we seek to analyze the affect of emotion in classifying DAs. Also, DAC in text usually involves extra information that can be benefitted from associated modalities. Below, we analyze some samples that require emotion aided and multi-modal reasoning. We exemplify using few instances from our proposed dataset in order to support our claim of DA often being expressed in a multi-modal way along with exploiting the emotional state of the speaker.

Role of Emotion. In Figure 2b, we present two instances from the dataset where the emotional state of the user seems beneficial in deciding the DA of an utterance. In the first example, the reference to the sad and dismal state of the speaker directs it to acknowledge the presence of the hearer. In the second case, the angry emotional state of the speaker forces her to disagree with people's opinion or suggestion involved in the conversation. The examples above illustrate the importance of having emotional information as emotions affect the communicative intention or DA of the speaker discussed above. The presence of emotion in our dataset caters the models with the ability to use additional information while reasoning about DA.

Role of Multi-modality. Figure 2a shows two cases where DA is articulated through incongruity between modalities. In the first instance, the facial modality implies anger or fury. Whereas the textual modality lacks any visible sign of displeasure, on the contrary it indicates an agreement. So, the textual claims does not validate the facial features. In the second case, the textual modality hints pure

agreement. Whereas the audio modality expresses a sarcastic appeal. In both these cases, there exists inconsistency between modalities, which acts as a strong indicator that multi-modal information is also important in providing additional cues for identifying DAs. The availability of complementary information across multiple modalities improves the model's ability to learn discriminatory patterns that are responsible for this complex process.

4 Proposed Methodology

This section describes the proposed multi-task, multi-modal approach followed by the implementation details.

4.1 Multi-modal Feature Extraction

Here, we discuss, the process of multi-modal feature extraction.

Textual Features. The transcriptions available for each video forms the source of the textual modality[2]. To extract textual features, pretrained GloVe (Pennington et al., 2014) embeddings of dimension 300 have been used to obtain representation of words as word vectors. The resultant word embeddings of each word are concatenated to obtain a final utterance representation. While it is indeed possible to use more advanced textual encoding techniques (for e.g., convolutional or recurrent neural network), we decided to use the same pre-trained extractive strategy as in the case of other modalities.

Audio Features. To elicit features from the audio, *openSMILE* (Eyben et al., 2010), an open source software has been used. The features obtained by openSMILE include maxima dispersion quotients (Kane and Gobl, 2013), glottal source parameters (Drugman et al., 2011), several low-level descriptors (LLD) such as voice intensity, voice quality (for eg., jitter and shimmer), MFCC, voiced/unvoiced segmented features (Drugman and Alwan, 2011), pitch and their statistics (for eg., root quadratic mean, mean etc.), 12 Mel-frequency coefficients etc. All the above features are then concatenated together to form a d_q = 256 dimensional representation for each window. The final audio representation of each utterance (A) is obtained by concatenating the obtained d_q for every window i.e., $A \in \mathbb{R}^{w \times d_q}$ where w represents total window segments.

Video Features. To elicit visual features for each of the f frames from the video of an utterance, we use a pool layer of an ImageNet (Deng et al., 2009), pretrained ResNet-152 (He et al., 2016) image classification model. Initially, each of the frames is preprocessed which includes resizing and normalizing. So, the visual representation of each utterance (F) is obtained by concatenating the obtained d_f = 4096 dimensional feature vector for every frame, i.e., $F \in \mathbb{R}^{f \times d_f}$ (Castro et al., 2019), (Illendula and Sheth, 2019), (Poria et al., 2017b), (Poria et al., 2017a).

4.2 Network Architecture

The proposed network consists of three main components : (i) *Modality Enocoders* (ME) which primarily takes as input the uni-modal features (extracted above) and produce as outputs the individual modality encodings, (ii) *Triplet Attention Subnetwork* (TAS) that encompasses self, inter-modal and inter-task attention and (iii) *classification layer* that contains outputs of both the tasks (DAC and ER).

4.2.1 Modality Encoders

In this section, we discuss how different modalities are encoded in the architectural framework.

Textual Modality. The obtained utterance representation (U) from the extracted textual features (discussed above) is then passed through three different Bi-directional LSTMs (Bi-LSTMs) (Hochreiter and Schmidhuber, 1997) to sequentially encode these representations into hidden states and learn different semantic dependency based features pertaining to different task, i.e., DAC and ER. One Bi-LSTM learns DAC features that are tuned in accordance with the emotion features. Second learns features for the ER task regulated by the learning of DA features. The third Bi-LSTM learns private features for the task of DAC which is not influenced by the features learnt from emotion.

$$\overrightarrow{h_i} = LSTM_{fd}(u_i, \overrightarrow{h}_{i-1}), \quad (1)$$

$$\overleftarrow{h_i} = LSTM_{bd}(u_i, \overleftarrow{h}_{i+1}). \quad (2)$$

For each of these word features, its corresponding forward and backward hidden states $\overrightarrow{h_i}$, $\overleftarrow{h_i}$, respectively, from the forward $LSTM_{fd}$ and the backward $LSTM_{bd}$ are concatenated to obtain a

[2]Original dataset with its video and transcript are downloaded from : https://github.com/SenticNet/MELD, https://sail.usc.edu/iemocap/iemocap_release.htm

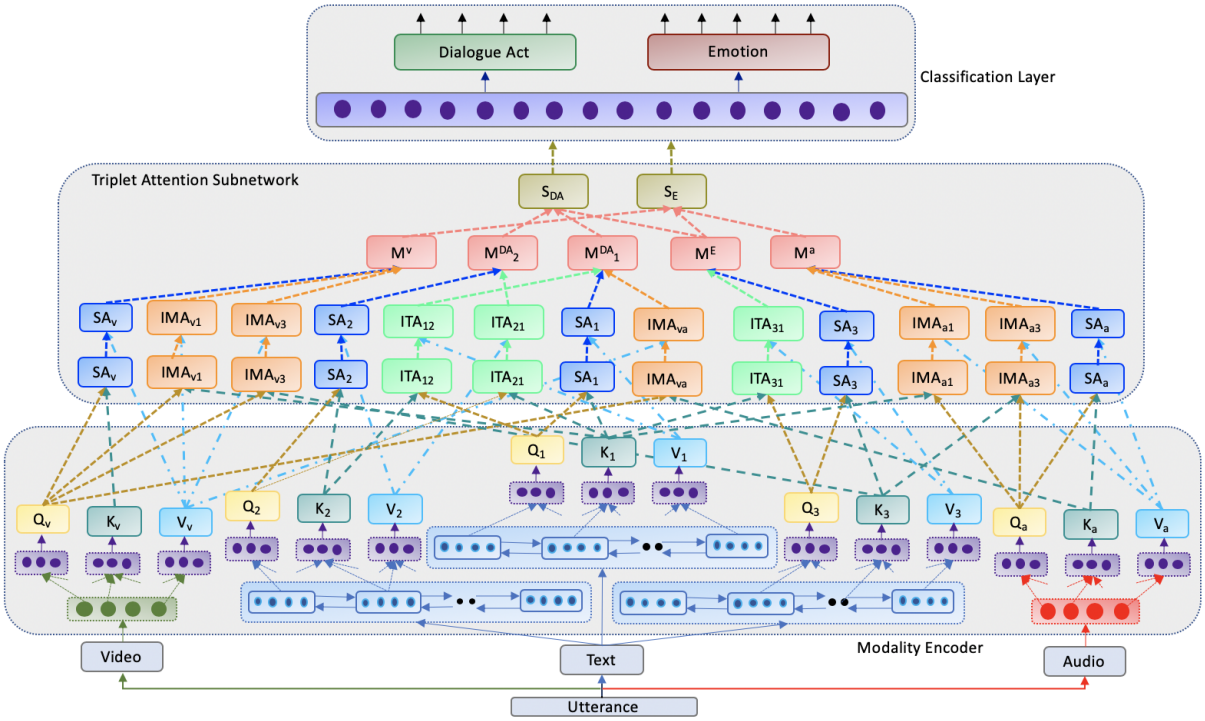


Figure 4: The architectural diagram of the proposed network. SA, IMA, ITA represent self, inter-modal and inter-task attentions, respectively.

single hidden state h_i. The complete hidden state matrix is obtained as,

$$H = [h_1, h_2, ..., h_n], \tag{3}$$

where $H \in \mathbb{R}^{n \times 2d}$. d represents the number of hidden units in each LSTM and n is the sequence length. Thus, the obtained three hidden state matrices correspond to three Bi-LSTMs, i.e., H_1, H_2, H_3. These representations are then passed through three fully connected layers, each of dimension say d_c to learn attention of different variants.

Audio and Video Modalities. The audio and video features (A and F) extracted are also passed through three fully connected layers, each of dimension say d_c, to learn attention of different variants.

4.2.2 Triplet Attention Subnetwork

We use a similar concept as in (Vaswani et al., 2017), where the authors proposed to compute attention as mapping a query and a set of key-value pairs to an output. The output is estimated as a weighted sum of the values, where the weight assigned to each value is calculated by a compatibility function of the query with its corresponding key. So, the representations obtained from each of the modality encoders above which are passed through three fully-connected layers each are termed as queries and keys of dimension $d_k = d_c$ and values of dimension $d_v = d_c$. We now have five triplets of (Q, K, V) as : (Q_1, K_1, V_1), (Q_2, K_2, V_2), (Q_3, K_3, V_3), (Q_a, K_a, V_a), (Q_v, K_v, V_v) where first three triplets are from the textual modality encoder (one each for DA_shared, DA_private and Emotion_shared)[3] followed by one from audio and video encoder each. These triplets are then used in different combinations to compute attention scores meant for specific purposes that includes self attention, inter-modal attention and inter-task attention.

Self Attention. We compute self attention (SA) for all these triplets by computing the matrix multiplication of all its corresponding queries to its corresponding keys.

$$SA_i = Q_i K_i^T \tag{4}$$

where $SA \in \mathbb{R}^{n \times n}$ for SA_1, SA_2, SA_3, $SA \in \mathbb{R}^{n \times w}$ for SA_a, $SA \in \mathbb{R}^{n \times f}$ for SA_v.

Inter-modal Attention. We compute inter-modal attention (IMA) amongst triplets of all the modalities for the multi-task by computing the matrix multiplication of combination of queries and keys of different modalities using Equation 4. In this manner, we obtain five IMA scores as $IMA_{v1} \in \mathbb{R}^{f \times n}$, $IMA_{v3} \in \mathbb{R}^{f \times n}$, $IMA_{a1} \in \mathbb{R}^{w \times n}$, $IMA_{a3} \in \mathbb{R}^{w \times n}$ and $IMA_{va} \in \mathbb{R}^{f \times w}$.

[3] Subscript 1, 2 and 3 represent DA_shared, DA_private and Emotion_shared representations, respectively.

This is done in order to identify significant contributions amongst different modalities to learn optimal features for an utterance.

Inter-task Attention. We compute inter-task attention (ITA) amongst triplets of different tasks from the textual modality by computing the matrix multiplication of combinations of queries and keys of different tasks using Equation 4. In this manner, we obtain three ITA scores as $ITA_{12} \in \mathbb{R}^{n \times n}$, $ITA_{21} \in \mathbb{R}^{n \times n}$ and $ITA_{31} \in \mathbb{R}^{n \times n}$. This is done in order to learn joint features of an utterance for identification of DAs and emotions.

Fusion of Attentions. We then obtain softmax of all these computed different attention scores to squash them in a range of [0,1] so that the ones having maximum contribution gets the highest probability values and vice-versa. We then compute the matrices of attention outputs for different tasks and modalities from the different attention scores as:

$$A = softmax(Q_i K_j^T) V_i \tag{5}$$

where $A \in \mathbb{R}^{n \times d_c}$. So, we obtain 13 different attention outputs from its corresponding attention scores which are $SA \in \mathbb{R}^{n \times d_c}$ for SA_1, SA_2, SA_3, $SA \in \mathbb{R}^{w \times d_c}$ for SA_a, $SA \in \mathbb{R}^{f \times d_c}$ for SA_v, $IMA_{v1} \in \mathbb{R}^{f \times d_c}$, $IMA_{v3} \in \mathbb{R}^{f \times d_c}$, $IMA_{a1} \in \mathbb{R}^{w \times d_c}$, $IMA_{a3} \in \mathbb{R}^{w \times d_c}$, $IMA_{va} \in \mathbb{R}^{f \times d_c}$, $ITA_{12} \in \mathbb{R}^{n \times d_c}$, $ITA_{21} \in \mathbb{R}^{n \times d_c}$ and $ITA_{31} \in \mathbb{R}^{n \times d_c}$.

Next, we obtain mean of different attention outputs in varying combinations to finally obtain representations for each of the modalities and tasks as M_1^{DA}, M_2^{DA}, M^E, M^v and M^a.

$$M_1^{DA} = mean(SA_1, IMA_{va}, ITA_{12}) \tag{6}$$

$$M_2^{DA} = mean(SA_2, ITA_{21}) \tag{7}$$

$$M^E = mean(SA_3, ITA_{31}) \tag{8}$$

$$M^v = mean(SA_v, IMA_{v1}, IMA_{v3}) \tag{9}$$

$$M^a = mean(SA_a, IMA_{a1}, IMA_{a3}) \tag{10}$$

where $M \in \mathbb{R}^{1 \times d_c}$. Next, we focus on learning appropriate weights to combine these representations to obtain final sentence representation for each of the tasks to be optimized jointly.

$$W_1 = M_1^{DA} * M_2^{DA} \tag{11}$$

$$W_2 = M_1^{DA} * M^E \tag{12}$$

	IEMOCAP		MELD	
	# Utterance	**# Dialogue**	**# Utterance**	**# Dialogue**
Train	7497	242	7489	831
Test	1879	60	2500	208

Table 2: Statistics of the train and test set of the EMOTyDA dataset from different sources

where $*$ represents dot product of two vectors. Finally, we obtain sentence representation (S) for each of the tasks as follows:

$$S_{DA} = M_1^{DA} + W_1 * M_2^{DA} + W_2 * M^E \tag{13}$$

$$S_E = M^E * M^v * M^a \tag{14}$$

4.2.3 Classification Layer

The output, i.e., sentence representation for each of the tasks (S_{DA} and S_E) from the TAS are connected to a fully-connected layer which in turn consists of the output neurons for both the tasks (DAC and ER). The errors computed from each of these channels are back-propagated jointly to the successive prior layers of the model in order to learn the joint features of both the tasks thereby, allowing them to benefit from the TAS layer. As the main aim of this study is to learn DA with the help of emotion, the performance of the DAC task also banks on the quality of features learned for the ER task with useful and better features assisting the collective learning process and vice-versa.

4.3 Implementation

EMOTyDA dataset was divided into two parts of 80% - 20% split for train and test set respectively. The statistics of the train and test set are shown in Table 2. For all the experiments conducted, same train and test sets were employed to allow a fair distinction between all approaches. For encoding the textual modality, a Bi-LSTM layer with 200 memory cells was used followed by a dropout rate of 0.1. Fully-connected layer of dimension 300 was used in all the subsequent layers. The first and the second channel contain 12 and 10 output neurons, respectively, for the DA and the emotion tags. *Categorical crossentropy* loss function is used in both the channels. A learning rate of 0.01 was found to be optimum. *Adam* optimizer was used in the final experimental setting. *All these values are selected after a thorough sensitivity analysis of the parameters.*

5 Results and Analysis

EMOTyDA contains dialogues pertaining to dyadic and multi-party speakers, so, we performed experi-

	Dataset											
	EMOTyDA:dyadic				EMOTyDA:multiparty				EMOTyDA			
	DA		DA + ER		DA		DA + ER		DA		DA + ER	
Modality	**Acc.**	**F1-score**	**Acc.**	**F1-score**	**Acc.**	**F1-score**	**Acc.**	**F1-score**	**Acc.**	**F1-score**	**Acc.**	**F1-score**
Text (T)	63.75	60.67	65.23	62.35	46.20	39.23	48.90	41.10	53.56	49.17	53.02	50.22
Audio (A)	32.06	24.95	35.42	38.92	25.76	19.45	26.58	21.01	27.13	23.09	28.65	24.87
Video (V)	35.94	29.71	36.88	30.34	27.23	20.26	28.12	21.03	30.16	26.85	32.09	27.73
T + A	65.43	60.67	**66.98**	**62.08**	47.17	40.30	**49.42**	**41.69**	54.12	50.00	**56.62**	**51.99**
A + V	38.59	34.98	40.07	36.00	27.91	22.76	28.95	23.89	32.09	28.86	33.76	29.13
T + V	67.12	64.14	**70.55**	**68.12**	49.80	41.90	**51.00**	**44.52**	57.31	53.20	**60.88**	**57.96**
T + A + V	66.35	62.30	**69.45**	**67.00**	49.02	41.00	**50.65**	**44.00**	56.77	52.09	**59.86**	**56.05**
T + V (emotional cue)	65.26	60.20	-	-	46.88	39.70	-	-	54.31	50.02	-	-

Table 3: Results of all the baselines and the proposed models in terms of accuracy and F1-score. All the reported results are statistically significant

Model	**EMOTyDA (DA + ER)**	
	Acc.	**F1-score**
Feature level (early fusion)	51.20	48.09
Hidden-state level (late fusion)	53.27	49.80
Hypothesis level	50.93	47.31
T + V (SA)	56.76	49.84
T + V (IMA)	56.62	52.79
T + V (ITA)	56.99	52.23
T + V (SA + IMA)	56.62	51.70
T + V (SA + ITA)	58.48	52.62
T + V (IMA + ITA)	57.74	52.85
T + V	**60.88**	**57.96**

Table 4: Results of various baseline models for the multi-task framework for the EMOTyDA dataset

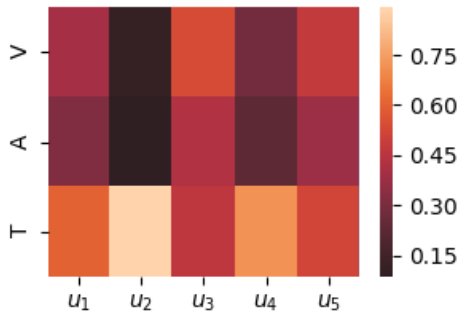


Figure 5: The visualization of the attention scores for 5 sample utterances for the tri-modal variant. *V*, *A* and *T* represent attention scores of video, audio and textual features, respectively. Sample utterance - u_1:"I am not in the least bit drunk.", u_2: "There's a lot of people looking for jobs.", u_3: "It was ridiculous. Completely ridiculous.", u_4: "You don't have to explain.", u_5: "No, Rachel doesn't want me to..."

ments segregating dyadic and multi-party conversations as well in addition to the whole dataset for the multi-task framework along with different modalities. Additionally, we also provide results of the multi-task framework with its varying combinations of different attentions applied to provide analysis on the effectiveness of each attention for the entire EMOTyDA dataset. Along with this, we also include results of some simple baselines such as feature level, hidden state level and hypothesis level concatenation. *It is to be noted that the purpose of the current work is to examine the effect of emotion while deciding the DA of an utterance from multiple modalities. We, therefore, do not focus on enhancements or analysis of the ER task and view it as an auxiliary task aiding the primary task, i.e., DAC*. In regards to this, the results and findings are reported with respect to only the DAC task and its different combinations.

Table 3 shows the results of all the various models. As visible, the textual modality provides the best results amongst the uni-modal variants. The addition of audio and visual features individually improves this uni-modal baseline. The combination of visual and textual features achieves the best score throughout all the combinations of the dataset. The tri-modal variant is not able to attain the best score supposedly because of suboptimal performance of the audio modality. Though it still improves the performance compared to all the uni-modal baselines. Figure 5 shows the heatmap visualization of the tri-modal variant to highlight the contributions of different modalities.

As is also evident from the results, the multi-task variant performs consistently well throughout all the experiments compared to its single task DAC variant. As a baseline, we also show that using emotion as a feature in the single task DAC counterpart doesn't outperform the proposed multi-task variant. This shows that the joint optimization of both these tasks boosts the performance of DAC. Table 4 shows the results of few simple baselines along with the ablation study of different attentions used in the proposed framework to highlight the importance and effectiveness of each of the attentions used for the whole EMOTyDA dataset. As seen from the table, the combinations of all three attention mechanisms, i.e., SA, IMA and ITA, yields the best results, thus, stressing the roles of incorporating across-task and across-modal relationships.

Figure 6 shows the visualization of the learned weights of different words for a sample utterance for the single task DAC as well as the multi-task model to highlight the importance of incorporat-

Utterance	True Label	MT(T+V)	ST (T+V)
She is not Larry's girl	dag	dag	s
I know, it was amazing! I mean, we totally nailed it, it was beautiful.	ag	ag	o
Then why is she still single?,New York is full of men.,Why hasn't she married? Probably a hundred people told her she's foolish, but she's waited.	o	s	q
God, I,feel so guilty about Ross.	ap	ap	s

Table 5: Sample utterances with its predicted labels for the best performing multi-task (MT) (T+V) model and its single task (ST) DAC variants; These examples show that ER as an auxiliary task helps DAC for better performance in MT.

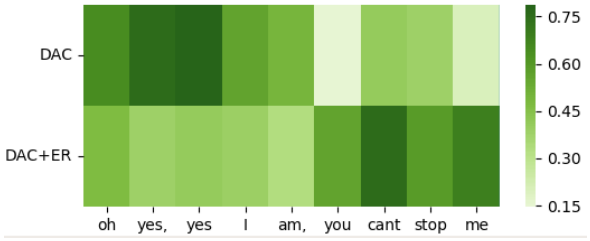


Figure 6: The visualization of the learned weights for an utterance - u_1: "Oh yes, yes I am, you can't stop me." for the best performing model (T+V), single task DAC (baseline) and multi-task DAC+ER (proposed) model

ing ER as an auxiliary task. The true DA label of the utterance in Figure 6 is *disagreement* with emotion as *anger*. With the multi-task approach, the attention is laid on appropriate disagreement bearing words whereas with single task, attention is learnt on agreement words such as *yes* which here has just been used in a sarcastic way to disagree. It is also observed that the experiments with dyadic conversations attain better results as compared to multi-party conversations. This is supposedly due to the constant change of speakers in multi-party conversations that misleads the classifier to learn suboptimal features, thus, stressing on the role of using speaker information as valuable cues for DAC.

Error Analysis. Plausible reasons behind the faults in the DA prediction are as follows : ***(i)*** **Skewed dataset** : The occurence of most of the tags in the proposed dataset is very less, i.e., the dataset is skewed as shown in Figure 1a. This consistently conforms with real time task-independent conversations where some tags occur less frequently as compared to others; ***(ii)*** **Composite and longer length utterance**: Most of the utterances in the dataset are longer in length and is also composite in nature encompassing diversified intentions in a single utterance. In such cases, it becomes difficult to learn features for discrete DAs; ***(iii)*** **Mis-classification of emotion labels**: Misclassification of the DAs can be attributed to the mis-classification of the emotions for that particular utterance. *Some examples for the same are shown in Table 5.*

6 Conclusion and Future Work

In this paper, we investigate the role of emotion and multi-modality in determining DAs of an utterance. To enable research with these aspects, we create a novel dataset, EMOTyDA, that contains emotion-rich videos of dialogues collected from various open-source datasets manually annotated with DAs. Consequently, we also propose an attention based (self, inter-modal, inter-task) multi-modal, multi-task framework for joint optimization of DAs and emotions. Results show that multi-modality and multi-tasking boosted the performance of DA identification compared to its unimodal and single task DAC variants. In future, conversation history, speaker information, fine-grained modality encodings can be incorporated to predict DA with more accuracy and precision.

Acknowledgments

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Lisa Feldman. Barrett, Michael Lewis, and Jeannette M. Haviland-Jones. 1993. *Handbook of emotions*. The Guilford Press.

Wauter Bosma and Elisabeth André. 2004. Exploiting emotions to disambiguate dialogue acts. In *Proceed-*

ings of the 9th international conference on Intelligent user interfaces, pages 85–92. ACM.

Kristy Elizabeth Boyer, Joseph F Grafsgaard, Eun Young Ha, Robert Phillips, and James C Lester. 2011. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1190–1199. Association for Computational Linguistics.

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. *Language resources and evaluation*, 42(4):335.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4619–4629.

Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255.

Thomas Drugman and Abeer Alwan. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Twelfth Annual Conference of the International Speech Communication Association*.

Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. 2011. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Pascale Fung, Dario Bertero, Peng Xu, Ji Ho Park, Chien-Sheng Wu, and Andrea Madotto. 2018. Empathetic dialog systems. In *The International Conference on Language Resources and Evaluation. European Language Resources Association*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Sergio Grau, Emilio Sanchis, Maria Jose Castro, and David Vilar. 2004. Dialogue act classification using a bayesian approach. In *9th Conference Speech and Computer*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Peter A Heeman and James F Allen. 1995. The trains 93 dialogues. Technical report, ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54.

Peter Lajos Ihasz and Victor Kryssanov. 2018. Emotions and intentions mediated with dialogue acts. In *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pages 125–130. IEEE.

Anurag Illendula and Amit Sheth. 2019. Multimodal emotion classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 439–449.

Neha Jain, Shishir Kumar, Amit Kumar, Pourya Shamsolmoali, and Masoumeh Zareapoor. 2018. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, 115:101–106.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. IEEE.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

John Kane and Christer Gobl. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011*, pages 169–176.

Nicole Novielli and Carlo Strapparava. 2013. The role of affect analysis in dialogue act identification. *IEEE Transactions on Affective Computing*, 4(4):439–451.

Daniel Ortega, Chia-Yu Li, Gisela Vallejo, Pavel Denisov, and Ngoc Thang Vu. 2019. Context-aware neural-based dialog act classification on automatically generated transcriptions. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7265–7269. IEEE.

Pinelopi Papalampidi, Elias Iosif, and Alexandros Potamianos. 2017. Dialogue act semantic representation and classification using recurrent neural networks. *SEMDIAL 2017 SaarDial*, page 104.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 973–982.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038. IEEE.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

Norbert Reithinger and Martin Klesen. 1997. Dialogue act classification using language models. In *Fifth European Conference on Speech Communication and Technology*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. A multilingual and multidomain study on dialog act recognition using character-level tokenization. *Information*, 10(3):94.

Tulika Saha, Saurabh Srivastava, Mauajama Firdaus, Sriparna Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Exploring machine learning and deep learning frameworks for task-oriented dialogue act classification. pages 1–8.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, Carol Van Ess-Dykema, et al. 1998. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5998–6008.

Daan Verbree, Rutger Rienks, and Dirk Heylen. 2006. Dialogue-act tagging using smart feature selection; results on multiple corpora. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 70–73. IEEE.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn W. Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Yue Yu, Siyao Peng, and Grace Hui Yang. 2019. Modeling long-range context for concurrent dialogue acts recognition. *arXiv preprint arXiv:1909.00521*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2236–2246.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2018. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

6.1 Appendix : Details of the Original Source Dataset

IEMOCAP : *Interactive Emotional Dyadic Motion Capture Database* (Busso et al., 2008) is a multi-modal ER dataset. It contains 151 videos of recorded dialogues, with 2 speakers per session for a total of 10 speakers in a two way conversation segmented into utterances amounting to a total of 302 videos across the dataset. Each utterance is annotated for the presence of 10 emotions namely *fear*, *sad*, *angry*, *frustrated*, *excited*, *surprised*, *disgust*, *happy*, *neutral* and *others*.

MELD : *Multi-modal EmotionLines Dataset* is also a multi-modal ER dataset derived from the Friends TV series, originally collected by (Poria et al., 2019). It contains 1433 dialogue conversations with multi-party speakers per dialogue amounting to a total of 13708 utterances across the dataset. Each utterance is annotated for the presence of 7 emotions namely *sadness*, *anger*, *fear*, *joy*, *surprise*, *disgust*, and *neutral*.