Meta-Learning based Deferred Optimisation for Sentiment and Emotion aware Multi-modal Dialogue Act Classification

Tulika Saha[*], **Aditya Prakash Patra**[†], **Sriparna Saha**[†], **Pushpak Bhattacharyya**[‡]
[*]University of Liverpool, United Kingdom
[†]Indian Institute of Technology Patna, India
[‡]Indian Institute of Technology Bombay, India
(sahatulika15,sriparna.saha,pushpakbh)@gmail.com

Abstract

Dialogue Act Classification (DAC) that determines the communicative intention of an utterance has been investigated widely over the years as a standalone task. But the emotional state of the speaker has a considerable effect on its pragmatic content. Sentiment as a human behavior is also closely related to emotion and one aids in the better understanding of the other. Thus, their role in identification of DAs needs to be explored. As a first step, we extend the newly released multi-modal *EMOTyDA* dataset to enclose sentiment tags for each utterance. In order to incorporate these multiple aspects, we propose a Dual Attention Mechanism (DAM) based multi-modal, multi-tasking conversational framework. The DAM module encompasses intra-modal and interactive inter-modal attentions with multiple loss optimization at various hierarchies to fuse multiple modalities efficiently and learn generalized features across all the tasks. Additionally, to counter the class-imbalance issue in dialogues, we introduce a 2-step Deferred Optimisation Schedule (DOS) that involves Meta-Net (MN) learning and deferred re-weighting where the former helps to learn an explicit weighting function from data automatically and the latter deploys a re-weighted multi-task loss with a smaller learning rate. Empirically, we establish that the joint optimisation of multi-modal DAC, SA and ER tasks along with the incorporation of 2-step DOS and MN learning produces better results compared to its different counterparts and outperforms state-of-the-art model.

1 Introduction

Dialogue Act Classification (DAC) constitutes an important means for understanding a speaker's communicative intention (for example, question, command, apology etc.) in any Dialogue System (Stolcke et al., 2000), (Papalampidi et al., 2017). Thus, DA seeks to analyze the pragmatics of a conversation instead of just its literal meaning. Authors of (Saha et al., 2020b) went a step ahead and established in a multi-modal setting (including text, audio and video) that a speaker's true communicative content is greatly influenced by its emotional state of mind (Barrett et al., 1993). Utterances such as "*Oh sure*" or "*Ya why not*" can be understood as "agreement" or "disagreement" (if implied sarcastically). However, the emotional state of the speaker might enclose cues giving it another definition altogether.

Sentiment and emotion are frequently viewed as two different entities (Do et al., 2019; Hossain and Muhammad, 2019; Majumder et al., 2019) etc., but are often interpreted in a similar way and are therefore used interchangeably due to their subjective character. But sentiment and emotion are not literally the same, but are strongly linked. For example, emotions such as *happy* and *joy* are inherently related to a *positive* sentiment. Thus, the speaker's emotion and sentiment are intertwined and one aids in better understanding of the other. As a result, information pertaining to emotion, as well as sentiment, provides a better comprehension of the speaker's state of mind. This strong relationship between emotion and sentiment drives us to incorporate the speaker's sentiment as well as its emotion while modeling DAs.

Additionally, we seek to address the class-imbalance issue for the task of DAC, as not all DAs are equally represented or are equally occurring in a conversation. When the training dataset has a high degree of class-imbalance, the testing criterion necessitates strong generalisation on less frequent classes (Neyshabur et al., 2017; Novak et al., 2018). To address this issue, a sample re-weighting approach is typically utilised (Sun et al., 2007; Lin et al., 2017; Kumar et al., 2010; Wang et al., 2017), which involves creating a weighting function that maps training loss to sample weight. Currently, employing this strategy requires manually pre-specifying the weighting function. However, this approach is not scalable in practice ow-

ing to the variations of an ideal weighing scheme based on the investigating task and training data at hand. In this paper, we leverage from the concept of meta-learning (Wu et al., 2018; Franceschi et al., 2018) to develop a method capable of learning an explicit weighting function from the data itself in an adaptable manner, named, *Meta-Net* (MN) learning. Simultaneously, we apply an effective training schedule (inspired by (Cao et al., 2019)) on top of MN Learning, namely, *two-step deferred optimization schedule* (2-step DOS). The 2-step DOS postpones or defers the re-weighting so that the classifier learns an initial representation while avoiding some of the complexities involved with re-weighting or re-sampling (incase of class-imbalance).

The contributions of this work are as follows : **(i)** We propose a *Dual Attention Mechanism* (DAM) based multi-task framework for multi-modal DAC, SA and ER in conversations. We leverage the information pertaining to emotional state and sentiment of the speaker to identify DAs; **(ii)** Additionally, we introduce a 2-step DOS that involves MN learning and deferred re-weighting to counter the class-imbalance issue for the task of DAC; **(iii)** In order to integrate these various facets, we extend the newly created dataset, EMOTyDA, to encompass annotations of the sentiment tags. We surmise that this extended characteristic of EMOTyDA will introduce novel sub-task for future investigation: sentiment and emotion aided DAC; **(iv)** We illustrate the gain in different measures that jointly optimizing these three tasks (DAC, SA and ER) using our proposed framework with the incorporation of 2-step DOS and MN learning produces better results compared to its different counterparts and state-of-the-art model.

2 Related Works

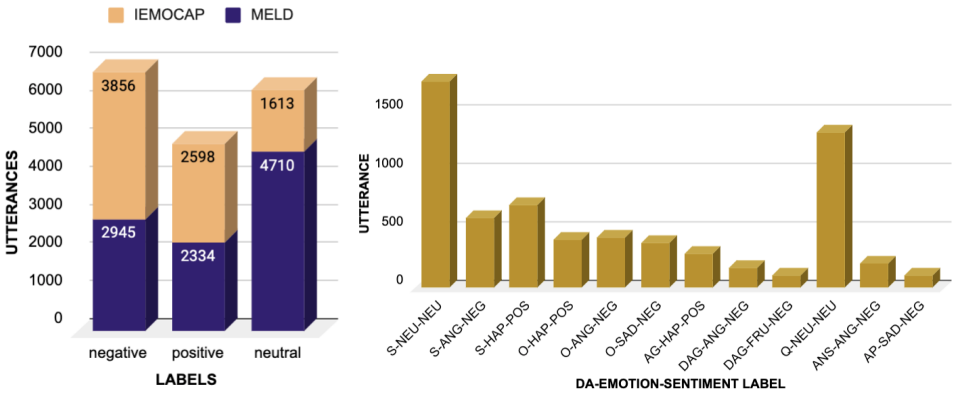
DAC, ER and SA are extensively explored linguistic tasks whose implications are observed in various dialogue system related research discussed below. With the success of Deep Learning (DL), DAC leveraged from it with several works proposed exploiting numerous DL concepts (Khanpour et al., 2016), (Kumar et al., 2018), (Khanpour et al., 2016) etc. However, all these works treated DAC as an independent problem without taking advantage of its correlation with other user behaviours such as emotion and sentiment. The idea of identifying speech acts in dialogues have also been extended for social media platforms such as Twitter also known as tweet acts (Saha et al., 2019, 2020c,d).

In (Cerisara et al., 2018b; Qin et al., 2020; Li et al., 2020), authors presented several DL based approaches to study the role of sentiment in identifying speech acts for a social media platform called Mastodon. In (Ihasz and Kryssanov, 2018), authors made an attempt to determine correlation between DAs and basic emotion tags for an *in-game* Japanese conversation. In (Saha et al., 2020b), authors introduced a large-scale, multi-modal conversational data annotated with DAs and emotions in order to establish that emotion indeed aided the task of DAC. However, they did not make use of sentiment of the speaker which is yet another crucial user behavior that can aid in understanding the DAs better. In (Saha et al., 2021, 2022), authors introduced the concept of studying speech acts in correlation with sentiment and emotion but it was meant for the social media communication in Twitter with no dialogic structure. Authors of (Saha et al., 2020f; Saha and Ananiadou, 2022; Saha et al., 2020e) proposed several correlated tasks in a dialogue system that leverages with the addition of sentiment and/or emotion in its learning process.

3 Dataset

The newly created multi-modal (i.e., text, audio and video), Emotion-DA Dataset: *EMOTyDA* (Saha et al., 2020b), consists of 1341 dyadic and multi-party conversations resulting in a total of 19,365 utterances and approximately 22 hours of recordings. In this dataset, utterances are annotated with 12 DA tags with the corresponding 10 emotion tags. The details of the DA and emotion tags are mentioned in the *appendix* below. So, this dataset is manually re-annotated for its related sentiment labels. EMOTyDA dataset is curated using conversations from MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008) datasets. In case of MELD, pre-annotated sentiment labels of the utterances were already existing. We chose to use the same sentiment labelling as released in the source dataset. However, the IEMOCAP dataset contains solely pre-annotated emotion tags without any sentiment labels[1]. Three annotators were hired for the task of sentiment annotation. They were asked to manually annotate the utterance by viewing the corresponding video and context to assign its senti-

[1]The extended version of the EMOTyDA dataset with its sentiment tags will be available in https://github.com/sahatulika15/EMOTyDA



(a) (b)

Figure 1: Statistics across the dataset : (a) Distribution of sentiment label, (b) Distribution of top-10 most highly occurring DA-Emotion-Sentiment labels.

ment label, namely *positive*, *negative* and *neutral*. We observed an inter-annotator score of 78% which can be considered reliable. Statistics of the dataset related to sentiment and relation between the different tasks are shown in Figure 1. Other statistics as well as the process of resolving disagreement amongst annotators are reported in the *appendix* below.

4 Proposed Methodology

The proposed approach and implementation details will be outlined in this section.

Problem Statement. For the multi-task set-up, let us consider a training set, $\{x_i, y_i, w_i, z_i\}_{i=1}^{N}$, where x_i is the i-th sample, y_i, w_i and z_i are the label vectors for DAC, SA and ER tasks, respectively and N is the number of training instances. $f(x, w)$ denotes the multi-task, multi-modal classifier, called the primary network (say) and w is its parameters. The task is to find the optimal parameter, w^*, by minimizing the multi-task training loss (combined loss from each of the three tasks), $1/N \sum_{i=1}^{N} L_i^{train}(w)$, where $L_i^{train}(w) = l(y_i, f(x_i, w))$.

4.1 Feature Extraction

The process of feature extraction for different modalities is discussed below.

• **Textual Features :** For extracting text based features of an utterance U having n_u number of words, the word embeddings of each of the words, $w_1, ..., w_u$, where $w_i \in \mathbb{R}^{d_u}$ and w_i's are obtained from pretrained GloVe (Pennington et al., 2014) embeddings, where $d_u = 300$. For an utterance U, each of these w_is belonging to the words of the utterance are concatenated to obtain a final textual sentence representation, i.e., $U \in \mathbb{R}^{n_u \times d_u}$.

• **Audio Features :** *OpenSMILE* (Eyben et al., 2010), an open source software has been used in order to extract features from the acoustic modality. Let n_a be the window segments for each of the audio with respect to an utterance. For each of the window segments, n_i, $d_a = 384$ dimension of features are obtained from the openSMILE software[2]. Each of these d_a dimensional features for n_a segments are concatenated to obtain a final audio representation for each of the utterances as $A \in \mathbb{R}^{n_a \times d_a}$.

• **Video Features :** To elicit visual features from the video of an utterance, containing n_v number of frames a pool layer of an ImageNet (Deng et al., 2009), pretrained ResNet-152 (He et al., 2016) image classification model has been used. For each of the frames, n_i, $d_v = 4096$ dimensional feature vector is obtained from the classification module. The final visual representation of each utterance (V) is acquired by concatenating each of the d_v vectors to a total of n_v, i.e., $V \in \mathbb{R}^{n_v \times d_v}$ (Castro et al., 2019), (Illendula and Sheth, 2019).

4.2 Network Architecture

The proposed network has three primary components : (i) *Modality Enocoders* (ME) which inputs the uni-modal features extracted above and output its respective modality encodings, (ii) *Dual Attention Mechanism* (DAM) comprising of *intra-modal* and *interactive inter-modal* attentions, (iii) *Classification Layer* containing output channels for

[2]We utilized the "The INTERSPEECH 2009 Emotion Challenge feature set" (IS09_emotion.conf) configuration file to extract the audio features

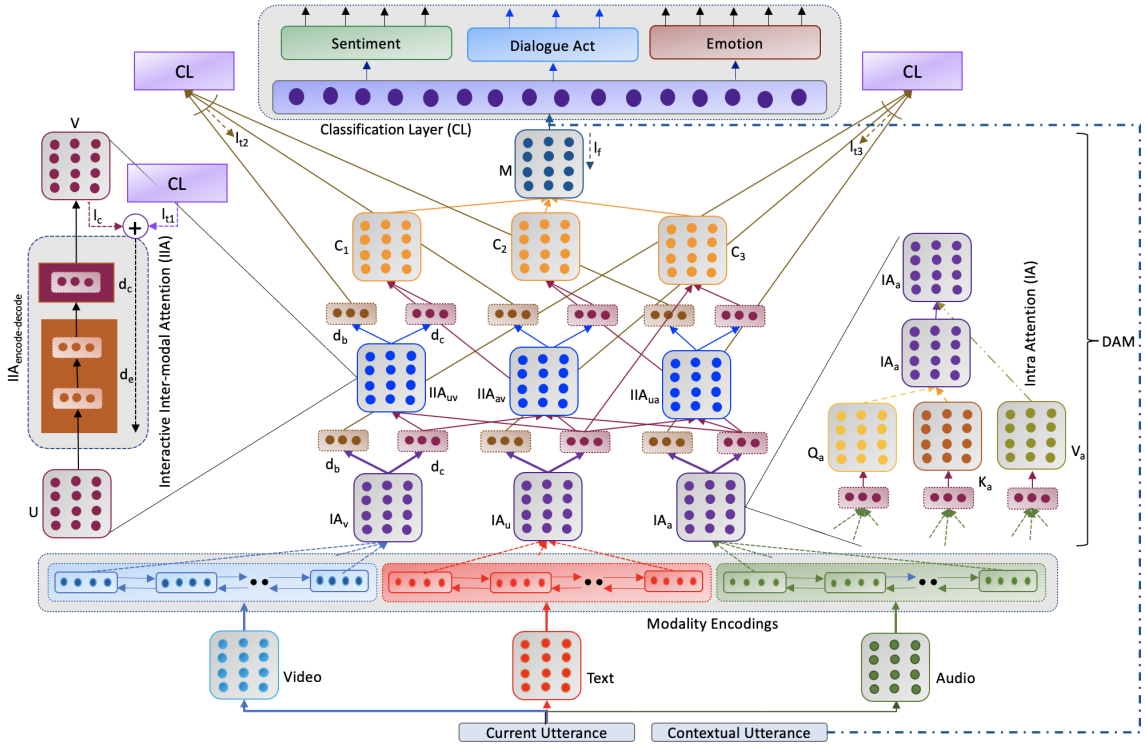


Figure 2: The architectural diagram of the proposed network

optimizing the three tasks (DAC, SA and ER) at different levels/hierarchies of the network to learn generalized representations.

Modality Encoders. Here, we detail how different modalities are encoded in the proposed architectural framework.

• **Text, Audio and Video Modalities :** The features U, A and V belonging to each of the modalities of an utterance (discussed above) are made to pass through three individual Bi-directional LSTMs (Bi-LSTMs) (Hochreiter and Schmidhuber, 1997). For textual modality (say), the corresponding representation of an utterance is shown as $H_u \in \mathbb{R}^{n_u \times 2d_l}$. Hidden units in each LSTM is represented as d_l and the sequence length is n_u. In this similar way, Bi-LSTMs are also applied to the features extracted from the audio and video modalities and finally a sentence representation of corresponding audio and video modality encodings as $H_a \in \mathbb{R}^{n_a \times 2d_l}$ and $H_v \in \mathbb{R}^{n_v \times 2d_l}$, respectively, is obtained.

Dual Attention Mechanism. One of the major challenges faced by any model employing multimodal inputs is to learn how to leverage the interactions amongst various modalities. Here, we introduce a Dual Attention Mechanism (DAM) for the joint optimization of DAC, SA and ER tasks. DAM primarily comprises of a series of attention mechanisms of varied types such as *intra-modal attention* (IA) and *interactive inter-modal attention* (IIA) that aim to learn complementary information from individual modalities as well as by interacting between two modalities.

• **Intra-modal Attention :** In order to understand how the current word and the preceding parts of the text are interdependent, we compute intra-modal attention (IA) for all of these modalities separately. So, we actually try to compute a final representation of the same sequence for each of these modalities by sort of relating different positions of that given sequence (Vaswani et al., 2017). The IA scores for each of the modalities are estimated as :

$$IA = softmax(Q_H K_H^T) V_H \quad (1)$$

where $IA \in \mathbb{R}^{n_u \times 2d_l}$ for IA_u, $IA \in \mathbb{R}^{n_a \times 2d_l}$ for IA_a, $IA \in \mathbb{R}^{n_v \times 2d_l}$ for IA_v.

Each of these matrices obtained from the individual modalities are then passed through individual dense layer of dimension, d_f (say). So, we obtain 3 different attention outputs from these modalities as $IA \in \mathbb{R}^{n_u \times d_f}$ for IA_u, $IA \in \mathbb{R}^{n_a \times d_f}$ for IA_a, $IA \in \mathbb{R}^{n_v \times d_f}$ for IA_v. Next, we obtain mean of these individual attention outputs to compute representations for each of these modalities in the same dimension as $IA \in \mathbb{R}^{1 \times d_f}$ for IA_u, $IA \in \mathbb{R}^{1 \times d_f}$ for IA_a, $IA \in \mathbb{R}^{1 \times d_f}$ for IA_v. These individual

representations are then passed through two separate dense layers of d_b and d_c (say) dimensions each. Thus, we obtain six different channels as $IA_{ub} \in \mathbb{R}^{1\times d_b}$, $IA_{uc} \in \mathbb{R}^{1\times d_c}$, $IA_{ab} \in \mathbb{R}^{1\times d_b}$, $IA_{ac} \in \mathbb{R}^{1\times d_c}$, $IA_{vb} \in \mathbb{R}^{1\times d_b}$ and $IA_{vc} \in \mathbb{R}^{1\times d_c}$.

• **Interactive Inter-modal Attention :** As stated above, one of the most challenging tasks for any multi-modal system is to successfully integrate inputs from various modalities. Individual modalities typically have discrete features, regardless of whether they contribute in the achievement of a common goal. For eg., in multi-modal DAC, the purpose of all the modalities i.e., text, audio and video is to predict the DA of a given utterance. The divergent characteristics from each modality alone is likely to provide an inconclusive scenario for deciding on a specific DA tag, reducing the model's ability to learn features efficiently. To counter this, we describe an interactive inter-modal attention (IIA) mechanism for learning a mutual interaction between two distinct modalities (in a way that the two modalities carry distinctive features of an utterance) serving a common goal. The IIA, thus, aims to encode feature representation of one modality (say text) and decode it into a feature representation of another modality (say video). In intuition, this concept is pretty similar to how an auto-encoder works. Like an auto-encoder aims to make the input and output conceptually as similar as possible. Analogously, the feature representations of two chosen modalities act as the input and the output, which are then meant to be conceptually aligned. In a sense, the IIA mechanism attempts to learn a vector that represents the combined representation of the two modalities involved which can thus, be further used in the network.

As seen in figure 2, the IIA network is implemented as a stacking of dense layers to deconstruct (encode) into lower dimension d_e and construct (decode) into higher dimension d_c of the input to the output. We take unique pairs of modality combination from IA_{uc}, IA_{ac}, IA_{vc} to form three unique pairs of input-output to feed to the IIA network resulting in $IIA_{ua} \in \mathbb{R}^{1\times d_c}$, $IIA_{uv} \in \mathbb{R}^{1\times d_c}$ and $IIA_{av} \in \mathbb{R}^{1\times d_c}$. In order to ensure that the resultant vector is as close to the output modality, the IIA vectors are conditionally trained using the cosine similarity loss, l_c where l_c is the maximizing function as for e.g., :

$$l_c = cos(IIA_{ua}, IA_a) \tag{2}$$

This applies to the remaining two IIA vectors as well. Also, while training this IIA network for each pair (say text-video), the encoded vector at the text side, i.e., IA_{uc} gets dual gradient of errors, one from the decoded IIA output at the video side, i.e., from IIA_{uv}, l_c and the other from the three task-oriented labels, l_{t1} of DAC, SA and ER. Both these errors are summed up, $(l_c + l_{t1})$ and back-propagated to the input side, i.e., IA_{uc} (shown in Figure 2). This is done so that the input side of the IIA network also adjusts itself to the desired task-specific features. To ensure, that output side of the IIA network (in this case IIA_{uv}) also learns features specific to the task, a gradient of error is also back-propagated to it for the three tasks at hand, l_{t2} (shown in Figure 2). This discussion also applies to other two IIA vectors as well.

• **Attention Fusion :** For each of these pairs, i.e., text-audio, text-video, audio-video, we obtain the corresponding IIA vectors along with the IA vectors of the encoded input vector. We concatenate each of these involved IA and IIA vectors:

$$C_1 = concat(IIA_{ua}, IA_u) \tag{3}$$
$$C_2 = concat(IIA_{av}, IA_a) \tag{4}$$
$$C_3 = concat(IIA_{uv}, IA_u) \tag{5}$$

where $C \in \mathbb{R}^{1\times 2*d_c}$ for each of C_1, C_2 and C_3. To get a final representation of the utterance, we take the mean of these three separate concatenated attention vectors.

$$M = mean(C_1, C_2, C_3) \tag{6}$$

Context. The context plays an essential role in deciding the DA of the current speaker (Liu et al., 2017). To incorporate the contextual relationship, previous utterance is encoded separately using a separate Bi-LSTM to model sentence level representation. The obtained contextual representation and the representation of the current utterance from the DAM module are concatenated to obtain a final representation.

Classification Layer. The final representation of an utterance obtained from the DAM module, is then passed through a dense layer and then shared across three channels of the proposed multi-task framework pertaining to the three tasks i.e., DAC, SA and ER. Each of these channels is accompanied by a *softmax* layer for the final classification. The gradient of errors, (l_f) received from each of these branches is back-propagated jointly to the preceding layers (shown in Figure 2). The three vectors,

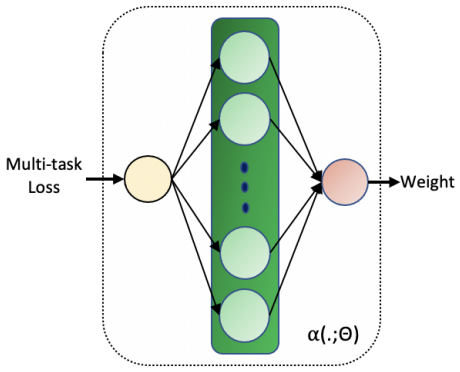


Figure 3: The architecture of the meta-net

IA_{ub}, IA_{ab} and IA_{vb}, obtained from the IA layer, are also subjected to the final classification layer separately, thus, receiving gradient of errors from the three task-oriented labels, l_{t3} (shown in Figure 2). In a way, these three vectors receive two gradients of errors to back-propagate, i.e., $(l_f + l_{t3})$. Similarly, the three vectors IIA_{uvb}, IIA_{avb} and IIA_{uab}, obtained from the IIA layer, are also subjected to the final classification layer separately, thus, receiving gradient of errors from the three task-oriented labels, l_{t2} as mentioned above. The intuition behind these multiple gradients of errors at some attention hierarchies is as DAC shares a lesser amount of correlation with SA and ER compared to SA and ER themselves, we impose a higher degree of strictness at various levels to learn useful features pertaining to the three tasks.

4.3 Meta-Net Learning

When the training data is biased, sample re-weighting based methods boost the efficiency of training by imposing weight on the i-th sample multi-task loss, $\alpha(L_i^{train}(w); \Theta)$, where $\alpha(l; \Theta)$ represents the weight net and Θ its parameters. The optimal w is calculated by minimizing the weighted multi-task loss as :

$$w^*(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \alpha(L_i^{train}(w); \Theta) L_i^{train}(w) \quad (7)$$

The MN-learning aims to exploit the idea of meta-learning to learn the hyper-parameters Θ automatically (inspired by (Shu et al., 2019)). For this, $\alpha(L_i^{train}(w); \Theta)$ is devised as a MLP network (shown in Figure 3). We refer to this weight net as Meta-Net. The input of MN is the multi-task loss and the output is a sigmoid function to squash the output in the interval of [0, 1]. We sample a small amount of unbiased data (focused on DAs, implying that sentiment and emotion might or might not be balanced) from the training set called the meta-data set, $\{x_i^{(meta)}, y_i^{(meta)}\}_{i=1}^{M}$ which depicts the meta-knowledge of DA ground-truth distribution, where M is the number of instances in meta-data set and $M << N$, the optimal Θ^* is obtained by minimizing the meta-loss as given below:

$$\Theta^* = \frac{1}{M}\sum_{i=1}^{M} L_i^{meta}(w^*(\Theta)) \quad (8)$$

So, the updating equation of the primary network (proposed framework discussed above) is devised by the current w^t along the descent direction of the multi-task loss in Eqn. 7 on a mini-batch training data as follows:

$$w^t(\Theta) = w^t - \gamma\frac{1}{n} \times \sum_{i=1}^{n} \alpha(L_i^{train}(w^t); \Theta) \bigtriangledown_w L_i^{train}(w) \quad (9)$$

where γ and n are step and mini-batch size, respectively. After receiving the feedback of the primary network, the parameter Θ is updated by moving the current Θ^t along the objective gradient of Eqn. 8 calculated on the meta-data as :

$$\Theta^{t+1} = \Theta^t - \beta\frac{1}{m}\sum_{i=1}^{m} L_i^{meta}(w^t(\Theta)) \quad (10)$$

where β is the step size. Thus, the updated Θ^{t+1} is utilized to alleviate the parameter w of the primary network as given below :

$$w^{t+1} = w^t - \gamma\frac{1}{n} \times \sum_{i=1}^{n} \alpha(L_i^{train}(w^t); \Theta^{t+1}) \bigtriangledown_w L_i^{train}(w) \quad (11)$$

4.4 Two-step Deferred Optimisation Schedule

Re-weighting and re-sampling are two well-known and successful procedures for dealing with imbalanced datasets because, as expected, they effectively bring the imbalanced training distribution closer to the uniform test distribution. The issues in applying these techniques are : (i) re-sampling the minority classes causes heavy over-fitting in DL based models (Cui et al., 2019) and (ii) when the minority class losses are weighted up, optimization can become difficult and unstable, especially when the classes are highly imbalanced (Huang et al., 2016). To counter this, we adopt a strategy similar to (Cao et al., 2019), known as deferred optimisation schedule. We call this two-step because at first

Table 1: Different hyper-parameter values used in the proposed approach

Hyper-parameter	Value
Bi-LSTM Memory Cells	100
Dense Layer (d_e, d_c, d_b)	100, 500, 300
Loss Function	Categorical Crossentropy
Learning Rate	0.01
Optimizer	Adam

we train the primary network with MN-learning before annealing the stochastic gradient descent learning rate, and then deploy a re-weighted multi-task loss with a smaller learning rate. Experimentally, the first step training induces a good initialization for the second step training. Since the multi-task loss is non-convex by nature and the learning rate for the second step is very small, it does not move the weights very far.

Implementation Details. 80% of the conversations of the EMOTyDA dataset were used as the train set and the remaining as the test set. The training set contains 14986 utterances resulting to 1073 dialogues whereas the test set comprises of 4379 utterances amounting to 268 dialogues. The three channels contain 12, 4 and 10 output neurons, for DA, sentiment and emotion tags, respectively. Different hyper-parameters and its value used in the proposed approach is listed in Table 1.

5 Results and Analysis

We carried out a number of experiments to assess the efficacy of the proposed method. Experiments were carried out for various combinations of multi-tasking with DAC as the crucial task, as well as for varying modalities, in addition to the single task DAC variation along with MN and DOS based learning. This was followed by experiments in a conversational framework and compared against single utterance classification.

Table 2 shows the results for all the baselines and the proposed models. As expected, the text modality gives the best results compared to the other two uni-modal variants (i.e., audio and video modality). However, as seen, the addition of these two non-verbal modalities improves this uni-modal textual baseline. Thus, stressing the role of considering multi-modal inputs for predicting DAs. The combination of text and video modalities (T+V) gives the best results compared to all other modality variants. The tri-modal variant does not achieve the best results due to the sub-optimal behavior of the acoustic modality. As evident in Table 2, the tri-task variant of the multi-task framework (i.e., DAC + SA + ER) consistently gave the best results throughout all the experiments, indicating that the presence of both sentiment and emotion benefits each other to comprehend the state of mind of the speaker better. All the reported results are statistically significant (Welch, 1947) as we have performed Welch's t-test at 5% significance level. As expected, in the bi-task variant, DAC+SA multi-task framework, shows little improvement in different metrics as opposed to DAC+ER multi-task framework compared to the single task DAC variant. This benefit is self-evident, as sentiment alone cannot always give a complete picture of the speaker's state of mind. For eg., a *negative* sentiment can arise due to various emotions such as *fear*, *disgust*, *sadness* etc.

In Table 3, we show experiments in different set-up by including contextual utterance along with the speaker utterance to predict the DAs. We observe that incorporating contextual relationship gave consistently better results for multi-task framework compared to single utterance classification. This observation is consistent with previous works. Additionally, we observe that the 2-step DOS involving MN learning and deferred re-weighting improves the performance of the DAC task considerably and consistently throughout all the multi-task variants. Intuitively, the incorporation of MN learning handles the extreme class-imbalance issue of the DAs effectively in a multi-task set-up. The addition of DOS on top of it further improves this issue indicating that the second-step of DOS starts from better features, adjusts the decision boundary and locally fine-tunes the features. All these observations are in conformity with the literature. We also compare our proposed approach with the recent state of the art models for different DAC and multi-modal models and the results for the same are reported in Table 5. As evident, the proposed network attained better results as compared to the state of the art models.

In Figure 4, we present a visualization of the learned weights of an utterance from the IA_u layer (as this layer contains word-wise attention scores). The true DA tag of this particular utterance is *disagreement*. The importance of disagreement bearing words are learnt well for the multi-task approach as opposed to the single-task DAC model where attention is on compliance bearing word such as *fine*. With DAC+SA, DAC+ER

			DAC		DAC + SA		DAC + ER		DAC + SA + ER	
Model	**MN**	**DOS**	**Acc.**	**F1-score**	**Acc.**	**F1-score**	**Acc.**	**F1-score**	**Acc.**	**F1-score**
Text (T)	×	×	51.27±1.08	48.63	51.92±1.02	49.06	52.65±0.75	49.76	53.00±0.80	50.46
Audio (A)	×	×	25.41±1.24	19.90	25.73±0.56	20.27	26.09±1.02	21.61	26.32±0.96	22.07
Video (V)	×	×	29.16±0.36	26.41	29.83±0.27	27.08	30.67±0.5	27.71	31.30±0.1	28.61
T+A	×	×	51.91±0.41	49.23	**52.57±0.29** †	**49.81** †	**53.36±1.41** †	**50.26** †	**54.61±1.2** †	**51.80** †
A+V	×	×	30.06±1.36	27.84	30.42±0.2	28.05	31.53±1.13	28.84	31.86±0.72	29.27
T+V	×	×	56.81±1.22	52.22	**57.27±1.08** †	**52.63** †	**58.12±0.51** †	**53.49** †	**58.56±0.54** †	**54.13** †
T+A+V	×	×	56.14±1.74	51.45	**56.81±2.31** †	**51.80** †	**57.34±1.28** †	**52.47** †	**57.81±1.42** †	**53.66** †
T+V (IA)	×	×	53.42±1.03	49.27	54.29±1.31	50.07	54.88±1.04	51.01	55.87±0.55	51.69
T+V (IIA)	×	×	52.63±1.3	48.91	53.77±1.05	49.65	54.06±0.61	50.33	54.63±0.16	50.60
T+V (single loss)	×	×	52.85±1.35	48.81	53.69±1.01	49.87	54.44±0.53	50.75	55.29±1.28	51.29
T+V (final concat attention)	×	×	54.21±0.56	49.72	55.09±0.36	50.35	55.82±1.20	51.06	56.31±0.67	52.05
T+V with Vanilla Re-weighting	×	×	56.83	52.66	57.76	52.90	58.49	53.82	58.91	54.56
T+V	✓	×	57.29	53.94	58.37	53.85	59.51	54.35	59.20	55.92
T+V	✓	✓	**58.72** †	**54.50** †	**59.96** †	**54.18** †	**60.02** †	**55.93** †	**61.72** †	**57.01** †

Table 2: Results of the proposed model (without context) and its different baseline in terms of accuracy and F1-score. † represents that the results are statistically significant

Model	**DAC + SA + ER (context)**	
	Acc.	**F1-score**
Text (T)	53.88±0.40	51.09
Audio (A)	26.61±0.39	22.19
Video (V)	31.75±0.62	28.94
T+A	55.46±1.27	52.25
A+V	32.35±1.21	29.74
T+V	**59.50±1.46** †	**54.86** †
T+A+V	58.73±1.02	54.08
T+V (IA)	56.82±1.52	52.22
T+V (IIA)	55.37±0.61	51.04
T+V (single loss)	56.62±1.23	51.74
T+V (final concat attention)	57.15±0.65	52.79

Table 3: Results of the proposed model considering context of the speaker utterance

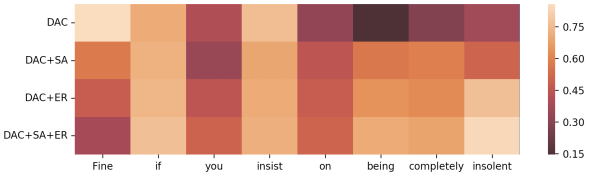


Figure 4: The visualization of the learned weights for an utterance from IA_u layer- u_1: "Fine, if you insist on being completely insolent." for the best performing model (T+V), single task DAC (baseline), multi-task DAC+SA, DAC+ER (baselines) and DAC+SA+ER (proposed) models

and DAC+SA+ER respectively, the degrees of importance of correct/incorrect words have increased/decreased gradually as enhanced information is learnt due to the effect of different tasks and its combinations. During a detailed analysis, it was observed that expressive DAs such as 'greeting", "acknowledge", "apology", "command", "agreement", "disagreement" are sensitive to the presence of sentiment and emotion etc. For e.g., utterance such as "*That's very amusing indeed*" was identified as "agreement" in the single task DAC model, but was correctly classified as "disagreement" in the proposed multi-task, DAC+SA+ER model as the sentiment and emotion of the utterance were "negative" and "angry", respectively, given the context that the speaker was disagreeing with the hearer in a sarcastic manner. It was also observed that for longer utterances comprising of composite sentences, sentiment and emotion of the speaker did play significant role in correctly identifying the DA tag. For eg., an utterance such as "*Hey, I'm, uh. I'm really sorry about what happened. I don't um- I mean what you can you do?*" was wrongly predicted as "opinion" in the single task DAC model but was predicted correctly as "apologize" in the proposed multi-task model given the "negative" and"sad" sentiment and emotion of the speaker, respectively, that it is simply trying to sympathize with the sufferer. It was also observed that "surprise" emotion gets marginally benefited with the addition of sentiment as there was no clear correlation of "surprise" with a definite sentiment state of the speaker. Other emotion categories such as *happy, anger, sad* which had direct correlations with the DA tags as shown in Figure 1b get benefited with the addition of sentiment tags.

Error Analysis. An in-depth investigation identified several possible explanations for why the proposed approach faltered which are as follows : **(i) Imbalanced dataset** : Most of the DA tags in the EMOTyDA dataset are less frequent than others, which make the dataset highly imbalanced. Due to their lesser instances, the model is unable to learn its representations correctly; **(ii) Composite utterances** : A number of utterances in the dataset are of composite nature with elongated span of words.

Utterance	True Label	DAC	DAC with (SA+ER)
Of course I did want to a little further up the coast you know get away from all the lights and people and everything. Is it midnight, do they always start at midnight? Is that what it is midnight? How you doing, huh? You okay? That's good.	q	o	o
You know you probably didn't know this, but back in high school, I had a, um, major crush on you.	s	ans	o
Oh that's a great reason. It's no reason at all.	dag	ag	dag
I know, I know, I'm such an idiot.	o	s	o
All right. All right. Calm yourself.	c	ag	c

Table 4: Examples with its predicted labels for the multi-task DAC+SA+ER (T+V) and its single task DAC variant

Model	Accuracy	F1-score
Feature level (early fusion) (Poria et al., 2015)	51.50%	48.49
Feature level (early fusion) + simple attention	52.34%	49.85
Hypothesis level fusion (Poria et al., 2016)	51.23%	47.72
JointDAS (Cerisara et al., 2018a)	52.03%	49.26
Hidden-state level (late fusion) (Saha et al., 2020a)	53.77%	50.06
Hidden-state level (late fusion) + simple attention	54.55%	50.19
SA+IMA : DAC+ER (Saha et al., 2020b)	56.62%	51.70
Proposed Approach (DAC+ER)	58.12%	53.49
Proposed Approach (DAC+SA+ER)	59.50%	54.86
Proposed Approach (DAC+SA+ER) MN+DOS	61.72%	57.01

Table 5: Comparison of the proposed approach with the recent state of the art models

Thus, a single utterance exhibits multiple notions of DAs making it challenging for classification models to learn features to discriminate amongst DAs; **(iii) Mis-identification and absence of sentiment-emotion tags** : In cases, where sentiment-emotion (and/or) tags were incorrectly identified, resulted in DAs also being wrongly classified. Also, instances where sentiment-emotion tags are *neutral*, the DAC task cannot really take advantage of these behaviors to enhance its learning. *Sample utterances for the error analysis are shown in Table 4.*

6 Conclusion and Future Works

In this paper, we study the role of sentiment and emotion while modelling the task of DAC. For this, we propose a Dual Attention Mechanism based multi-modal, multi-tasking framework for jointly optimizing DAC, SA and ER tasks. The DAM module employs intra-modal and interactive inter-modal attentions with multiple loss optimization at various hierarchies in order to fuse multiple modalities efficiently and learn generalized features across all the tasks. Additionally, to counter the class-imbalance issue in dialogues, we introduce a 2-step DOS that involves MN learning and deferred re-weighting where the former is an adaptive sample weighting strategy to automatically learn an explicit weighting function from data and the latter deploys a re-weighted multi-task loss with a smaller learning rate. Empirical results indicate that the joint optimisation of DAC, SA and ER tasks along with the incorporation of 2-step DOS and MN learning produces better results compared to its counterparts and outperforms SOTA model. In future, we would like to explore which other human behavior can aid the performance of DAC along with proposing other classification models encompassing speaker information and other DL concepts.

Acknowledgements

Author, Dr. Sriparna Saha, acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. 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 conducting this research.

References

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

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.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1565–1576.

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.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa Le. 2018a. Multi-task dialog act and sentiment recognition on mastodon. *arXiv preprint arXiv:1807.05013*.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018b. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 745–754.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9268–9277. Computer Vision Foundation / IEEE.

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.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118:272–299.

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.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. 2018. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1563–1572. PMLR.

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.

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

M Shamim Hossain and Ghulam Muhammad. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5375–5384. IEEE Computer Society.

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.

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. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3440–3447.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 1189–1197. Curran Associates, Inc.

Jingye Li, Hao Fei, and Donghong Ji. 2020. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 616–626.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. 2017. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5947–5956.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. 2018. Sensitivity and generalization in neural networks: an empirical study. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544.

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

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.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8665–8672.

Tulika Saha and Sophia Ananiadou. 2022. Emotion-aware and intent-controlled empathetic response generation using hierarchical transformer network. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2020a. Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation*, pages 1–13.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020b. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020c. A transformer based approach for identification of tweet acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Tulika Saha, Srivatsa Ramesh Jayashree, Sriparna Saha, and Pushpak Bhattacharyya. 2020d. Bert-caps: A transformer-based capsule network for tweet act classification. *IEEE Transactions on Computational Social Systems*, 7(5):1168–1179.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2019. Tweet act classification : A deep learning based classifier for recognizing speech acts in twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2020e. Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PloS one*, 15(7):e0235367.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2020f. Towards sentiment-aware multi-modal dialogue policy learning. *Cognitive Computation*, pages 1–15.

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Towards sentiment and emotion aided multi-modal speech act classification in Twitter. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5727–5737, Online. Association for Computational Linguistics.

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask multimodal ensemble model for sentiment- and emotion-aided tweet act classification. *IEEE Transactions on Computational Social Systems*, 9(2):508–517.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1917–1928.

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.

Yanmin Sun, Mohamed S. Kamel, Andrew K. C. Wong, and Yang Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.*, 40(12):3358–3378.

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.

Yixin Wang, Alp Kucukelbir, and David M. Blei. 2017. Robust probabilistic modeling with bayesian data reweighting. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3646–3655. PMLR.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Lijun Wu, Fei Tian, Yingce Xia, Yang Fan, Tao Qin, Jian-Huang Lai, and Tie-Yan Liu. 2018. Learning to teach with dynamic loss functions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6467–6478.

A Appendix

EMOTyDA Dataset. The 12 DA tags of the EMOTyDA dataset are namely, *Statement-Opinion* (o), *Greeting* (g), *Statement-Non-Opinion* (s), *Question* (q), *Apology* (ap), *Answer* (ans), *Command* (c), *Agreement* (ag), *Backchannel* (b), *Disagreement* (dag), *Acknowledge* (a) and *Others* (oth) with the 10 emotion tags, namely, *angry*, *fear*, *sad*, *excited*, *frustrated*, *disgust*, *surprised*, *happy*, *neutral* and *others*. The DAs and emotion labels distribution of the EMOTyDA dataset across the source datasets are shown in Figure 6. Distribution of sentiment labels of the EMOTyDA dataset across the source datasets are shown in Figure 1a. The 10 most highly occurring DA-Emotion-Sentiment labels in the EMOTyDA dataset is shown in 1b.

Sentiment Annotation. In case of disagreement between annotators, we utilized its corresponding emotion category to assign it, its related sentiment category. This was done because for e.g., emotions such as *excited* and *happy* are more likely to belong to the *positive* sentiment class whereas emotions such as *fear*, *sad*, *angry*, *frustrated* and *disgust* can be clubbed together to belong to the *negative* sentiment class. Similarly, *neutral* and *others* emotions can inherently belong to the *neutral* sentiment tags, respectively. For the *surprised* emotion tag, annotators were strictly asked to resolve disagreement amongst themselves by mutual agreement as an emotion of *surprise* can arise both because of *positive* as well as *negative* sentiments.

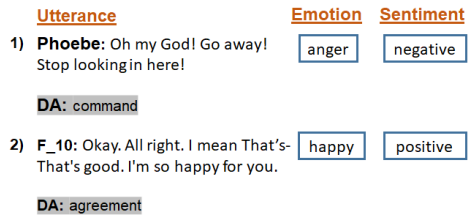
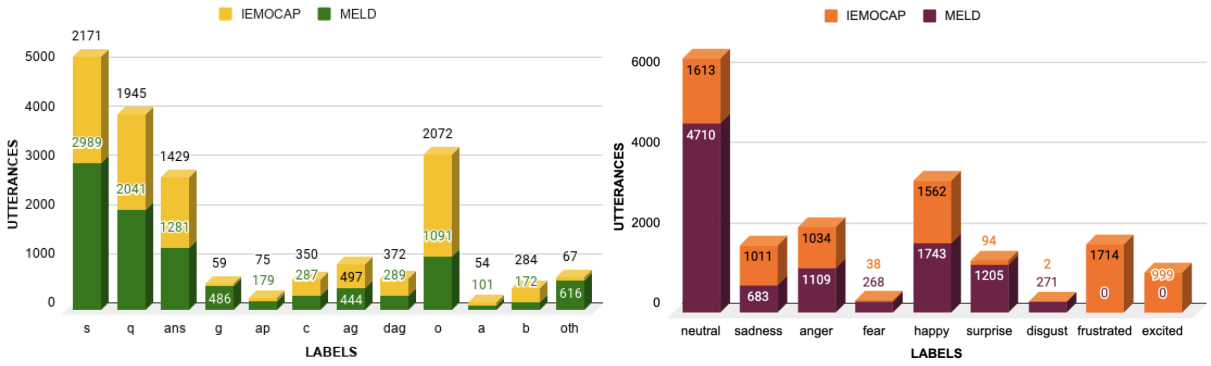


Figure 5: Importance of sentiment and emotion in DAC

Qualitative Aspect. Here, we investigate with some samples from the dataset that need sentiment and emotion needed reasoning for DAs. In Figure 5, we present two examples from the dataset and show how sentiment and emotional states of the speaker contribute in the identification of DAs. In the first instance, the commandment intent of the speaker is a result of her angry state of mind which in turn arises because of a negative sentiment. Similarly, in the second instance, the happier state of mind of the speaker largely directs the speaker to agree with the hearer which in turn can also be related to her positive sentiment. The above examples emphasize the importance of considering additional user behavior, such as sentiment and emotion, when reasoning about DAs. Thus, asserting the importance of resolving such synergy amongst DAC, SA, and ER.



(a) (b)

Figure 6: Statistics across the dataset : (a) Distribution of DA labels, (b) Distribution of emotion labels.