Action and Reaction go hand in hand! A Multi-modal Dialogue Act aided Sarcasm Identification

Mohit Tomar[1], Tulika Saha[2], Abhisek Tiwari[1] and Sriparna Saha[1]
[1]Indian Institute of Technology Patna, India
[2]University of Liverpool, United Kingdom
{mohitsinghtomar9797, sahatulika15, abhisektiwari2014, sriparna.saha}@gmail.com

Abstract

Sarcasm primarily involves saying something but "meaning the opposite" or "meaning something completely different" in order to convey a particular tone or mood. In both the above cases, the "meaning" is reflected by the communicative intention of the speaker, known as dialogue acts. In this paper, we seek to investigate a novel phenomenon of analyzing sarcasm in the context of dialogue acts with the hypothesis that the latter helps to understand the former better. Toward this aim, we extend the multi-modal *MUStARD* dataset to enclose dialogue acts for each dialogue. To demonstrate the utility of our hypothesis, we develop a dialogue act-aided multi-modal transformer network for sarcasm identification (*MM-SARDAC*), leveraging interrelation between these tasks. In addition, we introduce an order-infused, multi-modal infusion mechanism into our proposed model, which allows for a more intuitive combined modality representation by selectively focusing on relevant modalities in an ordered manner. Extensive empirical results indicate that dialogue act-aided sarcasm identification achieved better performance compared to performing sarcasm identification alone. The dataset and code are available at `https://github.com/mohit2b/MM-SARDAC`.

Keywords: Sarcasm Identification, Dialogue Act Classification (DAC), Multi-modality, Multi-tasking

1. Introduction

Sarcasm is an interesting phenomenon that creates a bitter, ironic impact on individuals, where the intended meaning is the "opposite of the literal meaning" of the speaker's utterance or "meaning something completely different" (Gibbs, 1986; Dews and Winner, 1995). Numerous studies have been conducted to detect sarcasm in textual settings (Joshi et al., 2017; Xiong et al., 2019; Srivastava et al., 2020), multi-modal settings (Wang et al., 2022; Liang et al., 2022) etc. Studies have also been carried out to perceive sarcasm in the realms of other affective behaviors such as sentiment and emotion of the speaker (Chauhan et al., 2020).

Identifying sarcasm is quite a challenging task as it requires discerning the underlying intended meaning or the pragmatics being conveyed, known as dialogue acts, rather than solely relying on the explicit utterance of the speaker. Multi-modal sarcasm detection has attracted attention in recent years (Castro et al., 2019). We must effectively fuse text, audio, and visual modalities to identify sarcasm in multi-modal settings. Different modalities play different roles in sarcasm identification. For example, if a speaker speaks angrily, we may not capture it from text modality, but a rise in one's tone can be captured through audio modality. Also, if the speaker is sarcastic, he/she may have a grin on his/her face, which can be captured from visual modality. Hence, audio and visual features can help identify sarcasm along with text modality. Also, different modalities have different importance when identifying sarcasm. For example, identifying

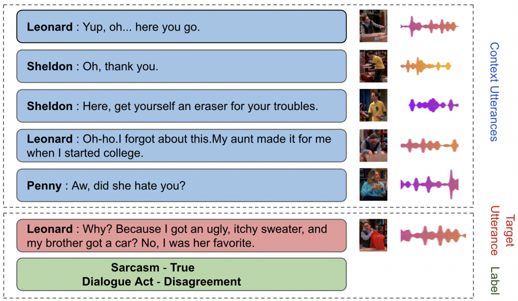

Figure 1: A conversational illustration to depict how sarcasm and DAs are related

changes in facial expression can be difficult compared to identifying meaning from text and audio modality. Hence, we fuse modalities differently to incorporate them according to their importance.

When humans engage in conversations, they often display specific communicative intentions known as Dialogue Acts (DAs), which can help in detecting the presence of sarcastic behavior in a speaker's utterance (Haverkate, 1990). Dialogue Act Classification is concerned with deciding the type of the speaker's utterance, i.e., communicative intention (question, statement, command, etc.). DAC is very important in discourse structure, as it supports intelligent dialogue systems, conversational speech transcription, and so on. For example, utterances such as "Okay, Sure" or "Ya right" can be considered as "sarcastic"- in case of- "disagreement" DA or "non-sarcastic"- in case of- "agreement" DA. Another such instance includes "praising

LREC-COLING 2024, pages 298–309
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

by criticizing" or vice-versa, "criticizing by praising." The corresponding dialogue act is needed to correctly characterize the presence of sarcasm. As seen in Figure 1, when *Penny* asks *Leonard* "*Aw, did she hate you*?", *Leonard* replied "*Why? Because I got an ugly, itchy sweater, and my brother got a car? No, I was her favourite*", it is observed that *Leonard's* communicative intent (DA) is to "*disagree*" with *Penny*, asserting the presence of sarcasm to mock the situation in *Leonard's* utterance. Therefore, detecting sarcasm is driven by understanding the pragmatics or the intended meaning of the speaker's utterance, i.e., DAs, and can be useful in simplifying the task of identifying sarcasm in conversations.

Over the years, sarcasm identification has mostly been investigated as a standalone task (Joshi et al., 2015; Babanejad et al., 2020). In this paper, we seek to investigate a novel phenomenon of aiding sarcasm with dialogue acts with the hypothesis that the latter helps to understand the former better. In this direction, we first extend the multi-modal MUStARD dataset (Castro et al., 2019) for sarcasm by annotating DAs for each dialogue. Our proposed approach, *MM-SARDAC*, involves extracting modality-specific features from the text, audio, and visual sources. These modalities are then integrated within a Pretrained Language Model (PLM) using an adapter block. We subsequently perform Dialogue Act-aided Sarcasm Identification and Sarcasm-aided Dialogue Act Classification, where the former is treated as our primary task and the latter as the auxiliary. Our findings demonstrate that the two tasks mutually enhance each other, as compared to when they are considered individually. By jointly considering the impact of DAs and incorporating multiple modalities in an ordered manner, our approach provides valuable insights into the task of sarcasm identification.

Contributions. We summarize our contributions as follows: (i) We propose a multi-modal framework for dialogue act-aided sarcasm identification and sarcasm-aided DAC in dialogues to study the role and impact of DAs for identifying sarcasm; (ii) The proposed model encompasses a *modality order* driven *modality fusion adapter* that fuses audio and visual signals using contextualized attention inside the BART model; (iii) We augment existing multi-modal sarcasm dialogue dataset to curate *Multi-modal Sarcasm-Dialogue Act Dataset,* $(MUStARD_2)$, having human annotated DA labels corresponding to each dialogue along with its pre-existing sarcasm labels; (iv) Empirical findings (both qualitative and quantitative) indicate the effectiveness of MM-SARDAC and DAs on sarcasm identification and shows its benefit over standalone task variants.

2. Related Works

The current work is mainly related to four research areas: sarcasm identification, DAC, multi-modal fusion, and parameter-efficient fine-tuning. In the following paragraphs, we have summarized the relevant works.

Sarcasm Identification. Castro et al. (2019) developed a benchmark multi-modal sarcasm identification dataset called MUStARD. Chauhan et al. (2020) developed a multi-task framework for detecting sarcasm with sentiment and emotion detection as auxiliary tasks. Liang et al. (2022) proposed a cross-modal graph-based model for identifying sarcastic utterances. Tomar et al. (2023) proposed to identify sarcasm in conversations by analyzing the importance of different modalities and incorporating them in a specific order.

Dialogue Act Classification. Malhotra et al. (2022) developed HOPE dataset that is used for DAC in counseling conversations. Raheja and Tetreault (2019) uses context-aware self-attention and hierarchical recurrent neural network to classify DAs. Wang et al. (2020) developed a neural generation model that is used to generate DAs and responses simultaneously. It maintains the meaning of multi-domain DAs, and in the generation process, it attends to DAs as needed. Ang et al. (2005) utilized lexical and prosodic knowledge sources for DA segmentation and DAC tasks using speech data in multi-party meetings. Saha et al. (2021a) and Saha et al. (2022) proposed to identify DAs in conversations with the help of the sentiment and emotion of the speaker in a multi-modal setting. The idea of studying speech acts has also been extended to social media conversations termed Tweet Acts (TAs) Saha et al. (2020a), Saha et al. (2019), Saha et al. (2020c) and have been further explored in the presence of emotion and sentiment of a tweeter in a multi-modal setting Saha et al. (2021b), Saha et al. (2021c).

Multi-Modality. Jaegle et al. (2021) developed a transformer-based neural network, Perceiver, that works with various modalities (text, audio, video, video + audio) and can scale to very large input dimensions. Alayrac et al. (2022) developed a visual language model that takes visual and textual inputs, returns textual output, and is able to perform a few shot learning on multi-modal tasks. Alayrac et al. (2020) used a self-supervised approach on how to combine text, audio, and visual modality and learning useful representation to help in downstream tasks. Suman et al. (2022) proposed a multi-modal system to predict personalities of different people.

Parameter-Efficient Fine-Tuning. Li and Liang (2021) developed a prefix tuning method that optimizes vectors that are prepended to key and value vectors in the transformer architecture. Houlsby et al. (2019) developed adapter modules that insert a small number of the trainable units inside the transformer layer for fine-tuning purposes. He et al. (2021) studied the connection between the transfer learning methods and proposed a new variant called Mix And Match adapter for fine-tuning purposes.

3. Dataset

To study the role of DAs in sarcasm detection in a multi-modal dialogue setting, we augment the existing dataset MUStARD with DA labels, along with its pre-annotated sarcasm labels, to introduce a new dataset called **Multi-modal Sarcasm-Dialogue Act Dataset, ($MUStARD_2$).**

3.1. Data Collection

To understand the role of DA on sarcasm, we select the benchmark and open-source multi-modal, conversational sarcasm dataset, MUStARD (Castro et al., 2019). The reason for its selection is that it is balanced in terms of both sarcastic and non-sarcastic labels. Also, due to its being multimodal, it was beneficial to study the role of DA in sarcasm in a multi-modal setting. To the best of our knowledge, we were unaware of any dataset that is annotated with both sarcasm and DAs in a multi-modal dialogic setting. Thus, the MUStARD dataset has been manually annotated with DAs to enhance the research in sarcasm identification. By incorporating DA annotations, we aim to provide a richer and more comprehensive resource for studying both sarcasm and DAs.

3.2. Data Annotation

For many years, the SWBD-DAMSL tag set having 42 DAs developed by Jurafsky (1997) has been widely used for DAC in task-independent dyadic conversations like SWBD (Godfrey et al., 1992). However, Saha et al. (2020b) proposed a collected taxonomy of the 12 most commonly occurring DAs influenced by the SWBD-DAMSL tag set, especially for smaller-sized conversational corpus. The motivation for using 12 DAs in EMOTyDA (as mentioned in Saha et al. (2020b)) was the lack of occurrence of 42 DAs in a smaller dialogue corpus. Our situation with the MUStARD dataset resonated with that of Saha et al. (2020b) as the MUStARD dataset is a very small dataset with just about 690 instances. Hence, we stuck with the 12 DAs and refrained from coming up with new tags to ease the course of study. Therefore, we use this taxonomy to annotate our $MUStARD_2$ dataset. The 12 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).

The current work annotates all 690 dialogues from the MUStARD dataset for its corresponding DAs. Three annotators qualified in English linguistics from the authors' affiliation were assigned the task of labeling each dialogue out of 12 possible DAs. The annotators were trained for the DA labeling task on an already existing benchmark dataset, EMOTyDA (Saha et al., 2020b), a multi-modal conversational dataset containing gold-standard labels for the DA tags. We chose this dataset to train the annotators because the 12 DA tags for the current work align with the EMOTyDA dataset. The annotators were initially provided with the subset of the EMOTyDA dataset to understand different examples of the DA tags. After a clear understanding of the tags, they were presented with another subset of the EMOTyDA dataset without the labels and were asked to annotate it. The annotated tags were compared with the existing gold-standard labels to identify discrepancies and further correct the annotators. Finally, all three annotators were presented with the MUSTARD dataset and were asked to annotate. They were asked to annotate these dialogues by viewing the video and transcript available without the information of pre-annotated sarcasm labels. This ensured the dataset wasn't biased to any specific sarcasm-DA labels.

The inter-annotator agreement score Cohen Kappa (Cohen, 1960) is 0.71, which indicates acceptable agreement. It is achieved in the first round of annotation of the MUSTARD dataset. This is reported based on the count that at least two annotators agreed on a particular DA tag, which was chosen as the final tag. The score stems from the fact that the annotators were initially trained on this task on a different dataset, but the annotators did better understand the task. The cases of disagreement were resolved with mutual discussion amongst the annotators and the primary author.

3.3. Multi-modal Sarcasm - Dialogue Act Dataset : $MUStARD_2$

The $MUStARD_2$ comprises 690 dyadic and multiparty conversations, resulting in a total of 2951 utterances. Each utterance contains three modalities: audio, text, and visual. We obtain the raw text, audio, and visual data from the MUStARD dataset and augment it with DA labels. The distribution of DA and sarcasm labels in the dataset is shown in Table 1.

Role of Dialogue Act. In Figure 2, we show two examples from the dataset where DA is useful in determining the sarcasm present in the conversa-

Dialogue Acts	Sarcasm	Non Sarcasm
Agreement	30	13
Answer	59	42
Statement Opinion	47	31
Disagreement	15	5
Question	38	54
Backchannel	7	19
Statement Non Opinion	19	8
Apology	3	8
Acknowledge	2	5
Command	2	8

Table 1: Distribution of dialogue acts and sarcasm labels

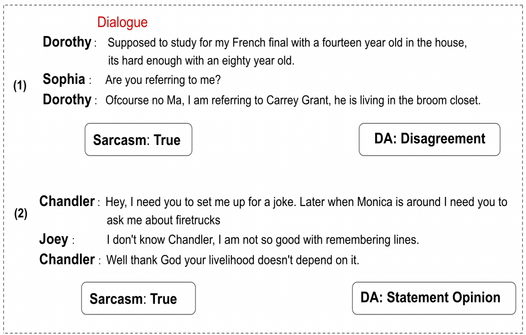

Figure 2: Importance of DAs in sarcasm detection

tion. In the first example, *Dorothy* is *disagreeing* and trying to refer to another person while indirectly indicating that she is referring to *Sophia* and that it is hard to study with her being around, hence being *sarcastic*. In the second example, *Chandler* is expressing an *opinion* about how *Joey's* livelihood doesn't depend upon remembering sentences and indirectly mocking him for not being good at memorizing things and hence being *sarcastic*. These examples show that DAs are useful in determining the presence of sarcasm and it presents the model with the ability to use additional information while reasoning about sarcasm.

4. Methodology

Our objective is to perform dialogue act-aided sarcasm identification and sarcasm-aided DAC in the view that the dialogue act helps identify sarcasm (and vice-versa through experiments). The proposed framework, *MM-SARDAC*, is illustrated in Figure 3. We describe how each of the modules works in the subsequent sections.

4.1. Multi-modal Feature Extraction

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

Textual Features. For extracting textual features, we use the BART-base (Lewis et al., 2019) model. It consists of BERT (Devlin et al., 2018) style encoder and GPT (Radford et al., 2018) style decoder. It generates the embedding for the textual input. For a given sentence S of j tokens, $\{s_1, s_2, ..., s_j\}$ it generates an embedding of dimension $S \in \mathbb{R}^{j\times 768}$.

Audio Features. For extracting audio features, we use Wave2Vec 2.0 (Baevski et al., 2020). Wave2Vec 2.0 is pretrained on LibriSpeech (Panayotov et al., 2015) and LibriVox data for Automatic Speech Recognition (ASR) task at a sampling rate of 16kHz. For sampling audio files, we use Librosa (McFee et al., 2015). The audio series is passed as an input to the Wave2Vec 2.0 model, and audio features are extracted from its last hidden state, $A \in \mathbb{R}^{a\times 768}$, where a is the audio segment length.

Video Features. We obtain video features from Castro *et al.* (Castro et al., 2019). The visual features are obtained corresponding to the video clips of the active speaker uttering the final utterances in which we want to identify the presence of sarcasm. Visual features of a video are obtained by extracting features from the pooling layer of a ResNet-152 (He et al., 2016) model for each of the frames. Thus, we obtain the final visual feature vector V, where $V \in \mathbb{R}^{v\times 2048}$, v is the number of frames in a video.

4.2. Network Architecture

The proposed network consists of three main components: (i) *Modality Encoding*, (i) *Modality Fusion Network* and (iii) *Central Network*.

Modality Encoding. We first obtain text representation by passing the whole dialogue as an input to BART, audio representation from Wave2Vec 2.0, and visual representation from the ResNet-152 model[1], respectively. All these three modality features are fused inside the BART at different layers in the following way.

Modality Fusion Network. We insert Modality Fusion Network as adapter unit (Houlsby et al., 2019) inside the BART encoder. The role of the adapter unit is to train only specific blocks called adapters while keeping the rest of the model frozen. Further, we obtain textual representation, T, from the lower transformer layers. For a particular layer, we generate query, key, and value vectors, Q, K, and V, respectively, from the textual representation, T, as shown in Equation 1. Here $T \in \mathbb{R}^{j\times d}$ $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ are learnable parameters, j is the sequence length of the model and d is the model dimension.

$$[Q, K, V] = [TW_q, TW_k, TW_v] \quad (1)$$

Let $M \in \mathbb{R}^{j\times d_m}$ denote the audio or video modality, where d_m is the modality dimension. We then

[1] Audio and Video representations are for the last utterance of a dialogue.

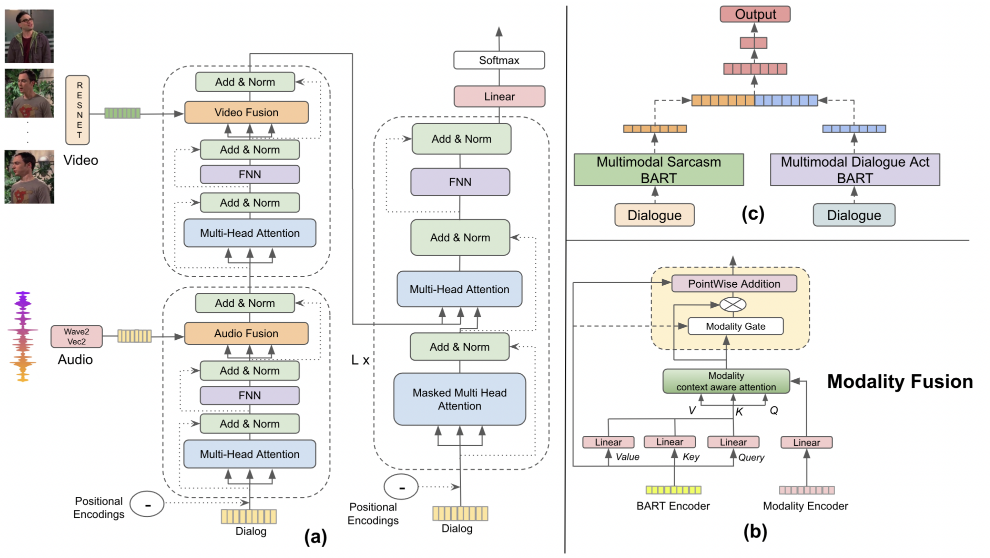

Figure 3: Architecture of (a) Multi-modal BART, (b) Modality Fusion Network. Here, Multi-modal BART consists of a Modality Fusion Network, indicated as audio and video fusion in the above figure. For fusion, it receives modality representation from lower layers of the BART encoder (for text), from Wave2Vec2.0 (for audio), and from ResNet-152 (for video). It is a component of *MM-SARDAC*, (c) Architecture of *MM-SARDAC* with the Central Network. Here Output can refer to *Sarcasm Identification* or *Dialogue Act Classification*

proceed to obtain modality contextualized key and value vectors.

$$\hat{K} = (1 - \lambda_k)K + (\lambda_k)(MU_k) \quad (2)$$

$$\hat{V} = (1 - \lambda_v)V + (\lambda_v)(MU_v) \quad (3)$$

λ_k and λ_v are learnable parameters given by the following equation :

$$\lambda_k = \sigma_1(KW_{k_1} + (MU_k)W_{k_2}) \quad (4)$$

$$\lambda_v = \sigma_1(KW_{v_1} + (MU_v)W_{v_2}) \quad (5)$$

The dimension of parameters are as follows:- λ_k and $\lambda_v \in \mathbb{R}^{j\times 1}, U_k$ and $U_v \in \mathbb{R}^{d_m \times d}$, W_{k_1}, W_{k_2}, W_{v_1} and $W_{v_2} \in \mathbb{R}^{d\times 1}$, and σ_1 means Sigmoid activation function.

Hence, we obtain the multi-modality infused key and value vectors. We then proceed to calculate the scaled dot product attention. In our case, we fuse audio and video modalities in different layers of the BART encoder (Kumar et al., 2022; Yang et al., 2019). Let m_1 and m_2 denote either audio or video modality. We first fuse modality m_1, and then we fuse modality m_2 with the text representation coming from the lower layers. For modality, m_1, we get contextualized key and value vectors as $\hat{K}_{t-m_1}$ and $\hat{V}_{t-m_1}$ from Equations 2 and 3. We calculate the scaled dot product attention with the text vector as:

$$C_{t-m_1} = \sigma_2\left(\frac{Q_t \hat{K}^T_{t-m_1}}{\sqrt{d_k}}\right)\hat{V}_{t-m_1} \quad (6)$$

The term d_k is the head dimension of a single head in multi-head attention, and σ_2 means Softmax activation function. The contextualized representation, C_{t-m_1}, is fused with textual representation from lower layers using a gated mechanism.

$$\hat{C}_{t-m_1} = T + g_{t-m_1} \odot C_{t-m_1} \quad (7)$$

$$g_{t-m_1} = [T \oplus C_{t-m_1}]W_1 + b_1 \quad (8)$$

where g_{t-m_1} represents gating mechanism, $\oplus$ means concatenation, $\odot$ means element wise multiplication, $W_1 \in \mathbb{R}^{2d\times d}$ and $b_1 \in \mathbb{R}^{d\times 1}$.

For fusing modality m_2 with the contextualized representation $\hat{C}_{t-m_1}$, we obtain a new set of query, key, and value vectors from $\hat{C}_{t-m_1}$ using Equation 1 and further obtain modality contextualized key and value vectors as $\hat{K}_{t-m_1-m_2}$ and $\hat{V}_{t-m_1-m_2}$, respectively from Equation 2 and 3. We then compute the scaled dot product attention as:

$$C_{t-m_1-m_2} = \sigma_2\left(\frac{Q_{t-m_1} \hat{K}^T_{t-m_1-m_2}}{\sqrt{d_k}}\right)\hat{V}_{t-m_1-m_2} \quad (9)$$

Next, the representations $C_{t-m_1-m_2}$ and $\hat{C}_{t-m_1}$ are fused through the gating mechanism.

$$\hat{C}_{t-m_1-m_2} = \hat{C}_{t-m_1} + g_{t-m_1-m_2} \odot C_{t-m_1-m_2} \quad (10)$$

$$g_{t-m_1-m_2} = [\hat{C}_{t-m_1} \oplus C_{t-m_1-m_2}]W_2 + b_2 \quad (11)$$

where $g_{t-m_1-m_2}$ is the gating mechanism, $W_2 \in \mathbb{R}^{2d \times d}$ and $b_2 \in \mathbb{R}^{d \times 1}$. The contextualized representation $\hat{C}_{t-m_1-m_2}$ is sent to the above layers for further processing.

Central Network. We perform the entire process (explained above) of modality encoding and fusion for each of the two tasks, i.e., DAC and sarcasm identification, using two different copies of the same architecture (see Figure 3). Firstly, we train the two models individually for sarcasm identification and DAC, respectively. Secondly, while performing classification in an aided manner, we freeze the individual parameters of the two models and obtain the representation from the last classification layer for both tasks, i.e., $hidden_{sar}$ and $hidden_{da}$. We then concatenate these two representations and then pass them to a linear layer followed by a non-linearity and classification layer for doing dialogue-act-aided sarcasm identification and sarcasm-aided dialogue act classification tasks.

$$shared_{task} = \sigma_3([hidden_{sar} \oplus hidden_{da}]W_3 + b_3) \tag{12}$$

From Equation 12, we obtain shared representations for both the tasks as $shared_{sar}$ and $shared_{da}$. We then pass these shared representations to the classification layer.

$$output_{sar} = shared_{sar}W_4 + b_4 \tag{13}$$

$$output_{da} = shared_{da}W_5 + b_5 \tag{14}$$

Here, σ_3 represents the ReLU activation function, and the dimensions of the parameters are as follows:- $W_3 \in \mathbb{R}^{2d \times d}$, $b_3 \in \mathbb{R}^{d \times 1}$, $W_4 \in \mathbb{R}^{d \times p}$, $b_4 \in \mathbb{R}^{p \times 1}$, $W_5 \in \mathbb{R}^{d \times q}$, $b_5 \in \mathbb{R}^{q \times 1}$, where p and q are the number of classes in sarcasm and DA tasks, respectively. In the *MM-SARDAC* model, only the Modality Fusion Network and classification layers are trainable, and the rest of the model is frozen. Hence, our model utilizes Parameter-Efficient Fine Tuning (PEFT).

4.3. Experimental Setup

We use a pre-trained BART-base (Lewis et al., 2019) language model for our task, which is implemented using hugging face library (Wolf et al., 2019) in PyTorch framework (Paszke et al., 2019). We run experiments on the extended dataset $MUStARD_2$. The dataset is split into training- 540, validation- 75, and test- 75 dialogic instances. The hyperparameters used are as follows: fusion of audio (5^{th} encoder layer), the fusion of video (6^{th} encoder layer), audio dimension (768), video dimension (2048), learning rate (1e-4), number of epochs (20), batch size (32), optimizer (Adam).

5. Results and Discussion

We use accuracy, weighted F1-score, precision, and recall measures to evaluate the performance of the proposed *MM-SARDAC* and compare it against several baselines and state-of-the-art (SOTA) models.

Comparison with the Baselines. We conduct several experiments to illustrate the performance of our hypothesis and model for sarcasm identification and DAC in standalone and aided settings, the role of different modalities, and the impact of modality order fusion in these scenarios.

Role of Aided Classification. Table 2 presents the results of *MM-SARDAC* for the task of sarcasm identification in both standalone and when it is aided by dialogue acts. As evident, when sarcasm is aided by dialogue acts the performance of sarcasm detection consistently over its standalone variant across all the combinations of modalities. Our proposed approach attained a performance improvement of **1.33**% and +**1.54**% in terms of accuracy and F1-score, respectively, on sarcasm detection as compared to its standalone counterpart (see row corresponding to *MM-SARDAC* (t-a-v) in Table 2). This indicates that dialogue acts indeed boost the performance of sarcasm detection, in line with our proposed hypothesis. Additionally, we also report results for the task of DAC to analyze its effect in the context of sarcasm. Table 3 shows the results of *MM-SARDAC* for the task of DAC in both standalone and aided settings. Interestingly, we observe that the performance of DAC also improves consistently when aided by sarcasm compared to its corresponding standalone variants. Our proposed model achieved a performance gain of +**5.34**% and +**5.84**% in terms of accuracy and F1-score, respectively, for the task of DAC as compared to its single task setting (see row corresponding to *MM-SARDAC* (t-a-v) in Table 3). Thus, the above observations support our hypothesis that DAs help in identifying sarcasm better and vice-versa.

Role of Modality. To analyze the importance of different modalities, we report an ablation study of our model in tri/bi/uni-modal settings. In the case of sarcasm identification (seen in Table 2), when the textual modality is dropped from the trimodal setting, the performance drops by -**10.67**% and -**10.67**% in terms of accuracy and F1-score, respectively (see row corresponding to *MM-SARDAC* (t-a-v) and Bimodal (a-v) in Table 2). While the performance drop was observed to be -**6.67**% and -**6.67**% for the exclusion of audio modality in terms of accuracy and F1-score, respectively (see row corresponding to *MM-SARDAC* (t-a-v) and Bimodal (t-v) in Table 2), and -**0.07**% for the exclusion of visual modality in terms of F1-score (see row corre-

Model Description		Dialogue Act Aided Sarcasm Setting				Standalone Sarcasm Setting			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
MM-SARDAC (t-a-v)		**0.8133**	**0.8133**	**0.8133**	**0.8133**	**0.8**	**0.8103**	**0.8**	**0.7979**
Trimodal	t-v-a	0.7333	0.7333	0.7333	0.7333	0.7333	0.7437	0.7333	0.7298
	a-t-v	0.72	0.7217	0.72	0.7197	0.6933	0.6933	0.6933	0.6932
	a-v-t	0.7466	0.7469	0.7466	0.7466	0.7333	0.7341	0.7333	0.7332
	v-t-a	0.64	0.6402	0.64	0.64	0.5466	0.5478	0.5466	0.5456
	v-a-t	0.6533	0.6555	0.6533	0.6525	0.64	0.6412	0.64	0.6396
Bimodal	t-a	**0.8133**	**0.8166**	**0.8133**	**0.8126**	**0.8133**	**0.8166**	**0.8133**	**0.8126**
	t-v	0.7466	0.7469	0.7466	0.7466	0.7466	0.7467	0.7466	0.7465
	a-t	0.6933	0.7210	0.6933	0.6846	0.7333	0.7451	0.7333	0.7306
	a-v	0.7066	0.7066	0.7066	0.7066	0.6933	0.6933	0.6933	0.6932
	v-t	0.6933	0.7061	0.6933	0.6893	0.6933	0.6962	0.6933	0.6916
	v-a	0.68	0.6802	0.68	0.6796	0.68	0.68	0.68	0.68
Unimodal	t	0.7333	0.7356	0.7333	0.7323	0.7066	0.7070	0.7066	0.7063
	a	0.6266	0.6291	0.6266	0.6238	0.64	0.6431	0.64	0.6387
	v	0.64	0.6402	0.64	0.64	0.6533	0.6555	0.6533	0.6525

Table 2: Results of all the baselines and *MM-SARDAC* for sarcasm detection in standalone and dialogue act aided settings. m_1-m_2-m_3 represents modality order to be m_1, m_2 and m_3.

Model Description		Sarcasm Aided DAC Setting				Standalone DAC Setting			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
MM-SARDAC (t-a-v)		**0.48**	**0.3868**	**0.48**	**0.4120**	**0.4266**	**0.3213**	**0.4266**	**0.3536**
Trimodal	t-v-a	0.3866	0.3193	0.3866	0.3431	0.4133	**0.3910**	0.4133	**0.3759**
	a-t-v	0.3733	0.2659	0.3733	0.2944	0.3866	0.2427	0.3866	0.2899
	a-v-t	0.3466	0.2597	0.3466	0.2831	0.2666	0.2834	0.2666	0.2559
	v-t-a	0.32	0.2218	0.32	0.2572	0.3333	0.2367	0.3333	0.2712
	v-a-t	0.36	0.2496	0.36	0.2899	0.3066	0.2239	0.3066	0.2521
Bimodal	t-a	**0.4533**	0.3136	**0.4533**	**0.3673**	0.3466	0.2602	0.3466	0.2936
	t-v	0.44	**0.3268**	0.44	0.3610	**0.4533**	0.2905	**0.4533**	0.3526
	a-t	0.44	0.3242	0.44	0.3631	0.4133	0.3134	0.4133	0.3345
	a-v	0.3066	0.3008	0.3066	0.3066	0.2133	0.2253	0.2133	0.2151
	v-t	0.32	0.2283	0.32	0.2660	0.32	0.2384	0.32	0.2721
	v-a	0.2666	0.2090	0.2666	0.2340	0.2933	0.2795	0.2933	0.2305
Unimodal	t	0.36	0.2830	0.36	0.3146	0.4	0.2560	0.4	0.2925
	a	0.36	0.2064	0.36	0.2407	0.3733	0.1847	0.3733	0.2203
	v	0.3333	0.2248	0.3333	0.2583	0.3333	0.2006	0.3333	0.2486

Table 3: Results of all the baselines and *MM-SARDAC* for DAC in standalone and sarcasm aided settings

Audio	Video	Model	Acc	Precision	Recall	F1
4	5	DA_Only	0.36	0.1419	0.36	0.2036
		Sarcasm_Only	0.8133	0.8166	0.8133	0.8126
		Sarcasm_Aided_ DA	0.36	0.1940	0.36	0.2214
		DA_Aided_Sarcasm	0.7866	0.7896	0.7866	0.7859
5	5	DA_Only	0.4533	0.3292	0.4533	0.3740
		Sarcasm_Only	0.7066	0.7127	0.7066	0.7050
		Sarcasm_Aided_ DA	0.4	0.3345	0.4	0.3575
		DA_Aided_Sarcasm	0.7733	0.7734	0.7733	0.7732
5	**6**	DA_Only	0.4266	0.3213	0.4266	0.3536
		Sarcasm_Only	0.8	0.8103	0.8	0.7979
		Sarcasm_Aided_ DA	0.48	0.3868	0.48	0.4120
		DA_Aided_Sarcasm	0.8133	0.8133	0.8133	0.8133
6	5	DA_Only	0.4133	0.3910	0.4133	0.3759
		Sarcasm_Only	0.7333	0.7437	0.7333	0.7298
		Sarcasm_Aided_ DA	0.3866	0.3193	0.3866	0.3431
		DA_Aided_Sarcasm	0.7333	0.7333	0.7333	0.7333

Table 4: Ablation Study - Fusion of audio and video modalities in different layers of the BART encoder. Here *DA* means Dialogue Act.

Model	Acc.	Prec.	Recall	F1
SVM* (Castro et al., 2019)	/	0.721	0.717	0.718
BERT† (Devlin et al., 2018) (only t)	0.68	0.6807	0.68	0.6798
BERT† (Devlin et al., 2018) (t-a-v)	0.7466	0.7467	0.7466	0.7465
MAG-BERT† (Rahman et al., 2020)	0.7333	0.7338	0.7333	0.7330
MISA† (Hazarika et al., 2020)	0.76	0.7717	0.76	0.7568
A-MTL* (Chauhan et al., 2020)	/	0.7709	0.7667	0.7657
QPM* (Liu et al., 2021)	/	0.7749	0.7761	0.7753
HKT* (Hasan et al., 2021)	0.7941	0.8035	0.7941	0.7925
HKT† (our data-split)	0.7361	0.7362	0.7361	0.7360
MM-SARDAC (Standalone Sarcasm)	0.8	0.8103	0.8	0.7979
MM-SARDAC (DA aided Sarcasm)	**0.8133**$^{\$}$	**0.8133**$^{\$}$	**0.8133**$^{\$}$	**0.8133**$^{\$}$

Table 5: Performance comparison of *MM-SARDAC* against SOTA models. Here, $\$$ indicates statistical significant findings ($p <$ 0.05 at 5% significance level). * indicates results reported from the paper. **†** indicates results reported by executing the code provided in the paper. Here *DA* means Dialogue Act.

sponding to *MM-SARDAC* (t-a-v) and Bimodal (t-a) in Table 2).

Similarly, in the case of DAC (see Table 3), we observe a performance drop of -**17.34**% and -**10.54**%, a drop of -**4**% and -**5.1**% and a drop of -**2.67**% and -**4.47**% corresponding to the exclusion of textual, audio and visual modality in terms of accuracy and F1-score, respectively (see row corresponding to *MM-SARDAC* (t-a-v), Bimodal (a-v), Bimodal (t-v) and Bimodal (t-a) in Table 3). From these observations, we can conclude that the importance of modality for the joint optimization of these tasks is as follows:- *text* > *audio* > *visual*.

Role of Modality Order Fusion. In order to understand the effectiveness of order for fusing the

Context
Howard: Hold that thought.
Howard: Hello?
Amy: How can I make this up to you?
Penny: The answer's in this puzzle box. Let's see if you can open it.
Bernadette: You could have at least warned him about the furniture.
Raj: That's what I said when we moved it.
Bernadette: You helped him?!

Current Utterance
Raj: No, Stuart picked out those throw pillows all on his own.

Models' Predictions

HKT	Sarcasm Standalone	MM-SARDAC	Sarcasm Gold label	DA Gold label
Non Sarcastic	Non Sarcastic	Sarcastic	Sarcastic	Answer

Figure 4: Performance of *MM-SARDAC* and other models on a common test case

Dialogue	Dialogue Act Ground Truth	Sarcasm Ground Truth	Sarcasm Standalone	MM-SARDAC
PERSON: Glad you guys could make it. LEONARD: Of course. PENNY: Wow, it looks really pretty in here. PERSON: Yeah, turns out half a dozen menorahs really sets a mood.	Agreement	True	False	True
SHELDON: It's not like I was invited to Richard Feynman's house and have anything better to do. AMY: Is this how the rest of the night's going to be? SHELDON: I don't know the future. SHELDON: Do you think there's a chance that an asteroid could hit the Earth, destroying Feynman's house and everyone in it? AMY: No, Sheldon. SHELDON: Then buckle up; you're in for a cranky night.	Others	False	False	True
PHOEBE: Definitely! RACHEL: Yeah, I'm pretty confident about that. That's what makes it so easy for me to be 80% happy for Monica and Chandler! RACHEL: It would be nice to have a little guarantee though.	Others	True	False	False

Figure 5: Qualitative analysis of predictions made by different models

modalities in our proposed approach, we present ablation results by varying the sequence of modalities. In the case of sarcasm identification (see Table 2), the best results were obtained when the modalities were fused in text –> audio –> visual sequence with a relative improvement of +**6.67**% and +**6.67**% in terms of accuracy and F1-score, respectively, in comparison to fusing the modalities in audio –> visual –> text sequence (see row corresponding to *MM-SARDAC* (t-a-v) and Trimodal (a-v-t) in Table 2). In the case of DAC (see Table 3), the text –> audio –> visual sequence provided the best results and a performance gain of +**9.34**% and +**6.89**% in terms of accuracy and F1-score, respectively, in comparison to the text –> visual –> audio sequence (see row corresponding to *MM-SARDAC* (t-a-v) and Trimodal (t-v-a) in Table 3). The improvement firmly supports that the proposed, *MM-SARDAC* performs effective information processing in the following order: *text (content) –> audio tone –> visual cues*. Additionally, in Table 4, we show the performance of dialogue act-aided sarcasm identification and sarcasm-aided dialogue act classification when we do a fusion of audio and video modalities at different layers of the BART encoder.

Comparison with the SOTA. We compare the proposed model, *MM-SARDAC's* performance with different state-of-the-art models for the task of sarcasm identification task as shown in Table 5. In the SOTA multi-modal BERT approach (see row-3), we fuse all three modalities inside BERT by concatenating them. As observed, the proposed *MM-SARDAC* surpasses all the SOTA approaches, indicating the efficacy of modality order fusion and central network for the task.

Qualitative Analysis. We analyze the performance of different models on a common test case shown in Figure 4 to comprehend their strengths and weaknesses. Our proposed model successfully identified an utterance as sarcastic, while the other models misinterpreted it as non-sarcastic. This superior performance can be attributed to the presence of DAs, which the model leverages to comprehend sarcasm effectively. Additionally, we report samples in Figure 5, providing information about cases where the model predicts correct and incorrect responses. Figure 6 illustrates the confusion arising in our proposed *MM-SARDAC* during testing.

We also analyze the case in Table 2 where we don't find improvement in sarcasm identification when aided by dialogue act (see Bimodal row). In this case, when we fuse audio or visual modality as the first modality, we find that the performance of dialogue act-aided sarcasm identification either remains the same (v-t and v-a) or drops (a-t) except for (a-v), where it increases. Also, in cases where text is fused first with other modalities, it remains the same (t-a, t-v). From these observations, we can say that in Bimodal cases when audio/visual modality is fused first, it doesn't exploit dialogue act features as we need textual features to support it because understanding from audio/visual modality alone is hard for the model when the model doesn't have text to augment it as a first modality.

We analyze the correlation between sarcasm and DA. During our dataset analysis, we encountered a strong correlation between Sarcasm-DA tags. For example, DA tags such as *disagreement*, *agreement*, *answer*, and *statement opinion* co-occur with the *sarcasm* tag. In contrast, tags such as *question*, *backchannel*, and *apology* co-occur more with the *non-sarcasm* tag. Table 1 in the paper shows the distribution of the sarcasm/non-sarcasm tags with the DA tags. Our hypothesis is established by the analysis reported in Figures 4 and 5. In Figure 4, the instance is sarcastic, but in the standalone sar-

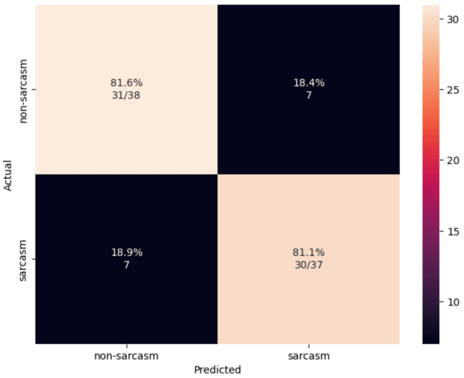

Figure 6: Confusion matrix of *MM-SARDAC* for sarcasm detection

casm model, the instance is wrongly predicted. But the inclusion of the DA task (in this case "answer") aids in the identification of sarcasm better.

6. Conclusion

In this work, we seek to investigate the role of DA and the order of multi-modality fusion in sarcasm identification task. As an attempt in this direction, we developed a multi-party conversational sarcasm identification dataset, $MUStARD_2$, that contains pre-existing sarcasm labels and newly annotated DA labels for each conversation. We propose a multi-modal framework for dialogue act-aided sarcasm identification and sarcasm-aided DAC in dialogues to study the role and impact of DAs for identifying sarcasm called *MM-SARDAC*. The extensive set of quantitative and qualitative experiments and the obtained improvements over state-of-the-art models firmly establish the efficacy of modeling dialogue act for sarcasm identification and vice versa. Sarcasm poses a highly abstract problem that necessitates a comprehensive contextual understanding for its identification. In the future, we aim to investigate the effectiveness of deep learning models infused with external knowledge to identify sarcastic utterances and generate a normalized explanation.

7. Ethical Consideration

While creating the dataset from the MUStARD dataset, we have not violated any copyright issues as the MUStARD dataset can be used for research purposes. We will make our code and dataset publicly available for research and reproducibility (when the paper is accepted). While annotating the dataset, annotators can be biased towards certain dialogue acts; thus, any biases in our dataset are not intentional.

8. References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37.

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1061. IEEE.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th international conference on computational linguistics*, pages 225–243.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shelly Dews and Ellen Winner. 1995. Muting the meaning a social function of irony. *Metaphor and Symbol*, 10(1):3–19.

Raymond W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1):3.

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

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12972–12980.

Henk Haverkate. 1990. A speech act analysis of irony. *Journal of pragmatics*, 14(1):77–109.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *www. dcs. shef. ac. uk/nlp/amities/files/bib/ics-tr-97-02. pdf*.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. *arXiv preprint arXiv:2203.06419*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777.

Yaochen Liu, Yazhou Zhang, Qiuchi Li, Benyou Wang, and Dawei Song. 2021. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 871–880.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 735–745.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics,* *speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2021a. Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation*, 13(2):277–289.

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

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020b. Towards emotion-aided multi-modal 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. IEEE.

Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Meta-learning based deferred optimisation for sentiment and emotion aware multi-modal dialogue act classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 978–990.

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. IEEE.

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021b. A multitask multimodal ensemble model for sentiment- and emotion-aided tweet act classification. *IEEE Transactions on Computational Social Systems*, 9(2):508–517.

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021c. 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.

Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava. 2020. A novel hierarchical bert architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97.

Chanchal Suman, Sriparna Saha, Aditya Gupta, Saurabh Kumar Pandey, and Pushpak Bhattacharyya. 2022. A multi-modal personality prediction system. *Knowledge-Based Systems*, 236:107715.

Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha. 2023. Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3926–3933.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. 2022. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8164–8175.

Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-domain dialogue acts and response co-generation. *arXiv preprint arXiv:2004.12363*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The world wide web conference*, pages 2115–2124.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 387–394.