When did you become so smart, oh wise one?! Sarcasm Explanation in Multi-modal Multi-party Dialogues

Anonymous ACL submission

Abstract

Indirect speech such as sarcasm achieves a constellation of discourse goals in human communication. While the indirectness of figurative language warrants speakers to achieve certain pragmatic goals, it is challenging for AI agents to comprehend such idiosyncrasies of human communication. Though sarcasm identification has been a well-explored topic in dialogue analysis, for conversational systems to truly grasp a conversation's innate meaning and generate appropriate responses, simply detecting sarcasm is not enough; it is vital to explain its underlying sarcastic connotation to capture its true essence. In this work, we study the discourse structure of sarcastic conversations and propose a novel task – ***Sarcasm Explanation in Dialogue*** (**SED**). Set in a multimodal and code-mixed setting, the task aims to generate natural language explanations of satirical conversations. To this end, we curate **WITS**, a new dataset to support our task. We propose **MAF** (Modality Aware Fusion), a multimodal context-aware attention and global information fusion module to capture multimodality and use it to benchmark WITS. The proposed attention module surpasses the traditional multimodal fusion baselines and reports the best performance on almost all metrics. Lastly, we carry out detailed analysis both quantitatively and qualitatively.

1 Introduction

The use of figurative language serves many communicative purposes and is a regular feature of both oral and written communication (Roberts and Kreuz, 1994). Predominantly used to induce humour, criticism, or mockery (Colston, 1997), paradoxical language is also used in concurrence with hyperbole to show surprise (Colston and Keller, 1998) as well as highlight the disparity between expectations and reality (Ivanko and Pexman, 2003). While the use and comprehension of sarcasm is a cognitively taxing process (Olkoniemi et al., 2016),

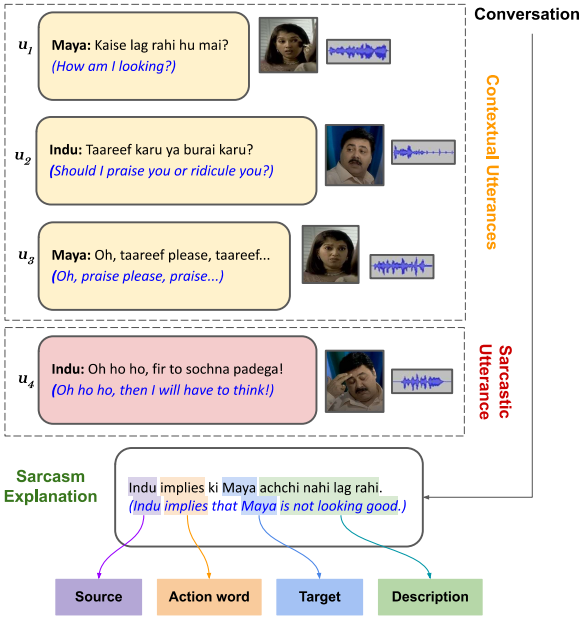


Figure 1: Sarcasm Explanation in Dialogues (SED). Given a sarcastic dialogue, the aim is to generate a natural language explanation for the sarcasm in it. *Blue text represents the English translation for the text.*

psychological evidence advocate that it positively correlates with the receiver's theory of mind (ToM) (Wellman, 2014), i.e., the capability to interpret and understand another person's state of mind. Thus, for NLP systems to emulate such anthropomorphic intelligent behavior, they must not only be potent enough to identify sarcasm but also possess the ability to comprehend it in its entirety. To this end, moving forward from sarcasm identification, we propose the novel task of ***Sarcasm Explanation in Dialogue*** (**SED**).

For dialogue agents, understanding sarcasm is even more crucial as there is a need to normalize its sarcastic undertone and deliver appropriate responses. Conversations interspersed with sarcastic statements often use contrastive language to convey the opposite of what is being said. In a real-world setting, understanding sarcasm goes beyond negating a dialogue's language and involves the acute

comprehension of audio-visual cues. Additionally, due to the presence of essential temporal, contextual, and speaker-dependent information, sarcasm understanding in conversation manifests as a challenging problem. Consequently, many studies in the domain of dialogue systems have investigated sarcasm from textual, multimodal, and conversational standpoints (Ghosh et al., 2018; Castro et al., 2019; Oraby et al., 2017; Bedi et al., 2021). However, baring some exceptions (Mishra et al., 2019; Dubey et al., 2019; Chakrabarty et al., 2020), research on figurative language has focused predominantly on its identification rather than its comprehension and normalization. This paper addresses this gap by attempting to generate natural language explanations of satirical dialogues.

To illustrate the proposed problem statement, we show an example in Figure 1. It contains a dyadic conversation of four utterances $\langle u_1, u_2, u_3, u_4 \rangle$, where the last utterance (u_4) is a sarcastic remark. Note that in this example, although the opposite of what is being said is, *"I don't have to think about it,"* it is not what the speaker means; thus, it enforces our hypothesis that sarcasm explanation goes beyond simply negating the dialogue's language. The discourse is also accompanied by ancillary audio-visual markers of satire such as an ironical intonation of the pitch, a blank face, or roll of the eyes. Thus, conglomerating the conversation history, multimodal signals, and speaker information, SED aims to generate a coherent and cohesive natural language explanation associated with these sarcastic dialogues.

For the task at hand, we extend MASAC (Bedi et al., 2021) – a sarcasm detection dataset for code-mixed conversations – by augmenting it with natural language explanations for each sarcastic utterance. We name the dataset WITS[1]. The dataset contains a compilation of sarcastic dialogues from a popular Indian TV show. Along with the textual transcripts of the conversations, the dataset also contains multimodal signals of audio and video.

We experiment with unimodal as well as multimodal models to benchmark WITS. Text, being the driving force of the explanations, is given the primary importance, and thus, we compare a number of established text-to-text systems on WITS. To incorporate multimodal information, we posit a unique fusion scheme of *Multimodal Context-Aware Attention* (MCA2). Inspired by Yang et al. (2019), MCA2 facilitates deep semantic interaction between the multimodal signals and textual representations by conditioning the key and value vectors with audio-visual information and then performing dot product attention with these modified vectors. The generated audio and video information-informed textual representations are then combined using the *Global Information Fusion Mechanism* (GIF). The gating mechanism of GIF allows for the selective inclusion of information relevant to the satirical language and also prohibits any multimodal noise from seeping into the model. We further propose MAF (*Modality Aware Fusion*) module where the aforementioned mechanisms are introduced in the Generative Pretrained Models (GPLMs) as adapter modules. Our fusion strategy outperforms the text-based baselines and the traditional multimodal fusion schemes in terms of multiple text-generation metrics. Finally, we conduct a comprehensive quantitative and qualitative analysis of the generated explanations.

In a nutshell, our contributions are four fold:

- We propose Sarcasm Explanation in Dialogue (SED), **a novel task** aiming to generate a natural language explanation for a given sarcastic dialogue to elucidate the intended irony.
- We extend an existing sarcastic dialogue dataset, to curate WITS, **a novel dataset** containing human annotated gold standard explanations.
- We **benchmark our dataset** using MAF-TAV$_{\text{B}}$ and MAF-TAV$_{\text{M}}$ variants of BART and mBART, respectively that incorporates the audio-visual cues using a unique context aware attention mechanism.
- We carry out **extensive quantitative and qualitative analysis** along with human evaluation to assess the quality of the generated explanations.

Reproducibility The source codes and the dataset, along with the execution instructions, are uploaded with the manuscript.

2 Related Work

Sarcasm and Text: Joshi et al. (2017) presented a well-compiled survey on computational sarcasm where the authors expanded on the relevant datasets, trends, and issues for automatic sarcasm identification. Early work in sarcasm detection dealt with standalone text inputs like tweets and reviews (Kreuz and Caucci, 2007; Tsur et al., 2010; Joshi et al., 2015; Peled and Reichart, 2017). These initial works mostly focused on the use of linguistic

[1] WITS: "Why Is This Sarcastic"

and lexical features to spot the markers of sarcasm (Kreuz and Caucci, 2007; Tsur et al., 2010). More recently, attention-based architectures are proposed to harness the inter- and intra-sentence relationships in texts for efficient sarcasm identification (Tay et al., 2018; Xiong et al., 2019; Srivastava et al., 2020). Analysis of figurative language has also been extensively explored in conversational AI setting. Ghosh et al. (2017) utilised attention-based RNNs to identify sarcasm in the presence of context. Two separate LSTMs-with-attention were trained for the two inputs (sentence and context) and their hidden representations were combined during the prediction.

The study of sarcasm identification has also spanned beyond the English language. Bharti et al. (2017) collected a Hindi corpus of 2000 sarcastic tweets and employed rule-based approaches to detect sarcasm. Swami et al. (2018) curated a dataset of 5000 satirical Hindi-English code-mixed tweets and used n-gram feature vectors with various ML models for sarcasm detection. Other notable studies include Arabic (Abu Farha and Magdy, 2020), Spanish (Ortega-Bueno et al., 2019), and Italian (Cignarella et al., 2018) languages.

Sarcasm and Multimodality: In the conversational setting, MUStARD, a multimodal, multi-speaker dataset compiled by Castro et al. (2019) is considered the benchmark for multimodal sarcasm identification. Chauhan et al. (2020) leveraged the intrinsic interdependency between emotions and sarcasm and devised a multi-task framework for multimodal sarcasm detection. Currently, Hasan et al. (2021) performed the best on this dataset with their humour knowledge enriched transformer model. Recently, Bedi et al. (2021) proposed a code-mixed multi-party dialogue dataset, MASAC, for sarcasm and humor detection. In the bimodal setting, sarcasm identification with tweets containing images has also been well explored (Cai et al., 2019; Xu et al., 2020; Pan et al., 2020) .

Beyond Sarcasm Identification: While studies in computational sarcasm have predominantly focused on sarcasm identification, some forays have been made into other domains of figurative language analysis. Dubey et al. (2019) initiated the work of converting sarcastic utterances into their non-sarcastic interpretations using deep learning. In another direction, Mishra et al. (2019) devised a modular unsupervised technique for sarcasm generation by introducing context incongruity through fact removal and incongruous phrase insertion. Following this, Chakrabarty et al. (2020) proposed a retrieve-and-edit-based unsupervised framework for sarcasm generation. The proposed model leverages the valence reversal and semantic incongruity to generate sarcastic sentences from their non-sarcastic counterparts.

A lot of effort has gone into detecting sarcasm, but very little, if any, effort has gone into *explaining the irony*. This paper attempts to fill the gap by proposing a new problem definition and a supporting dataset.

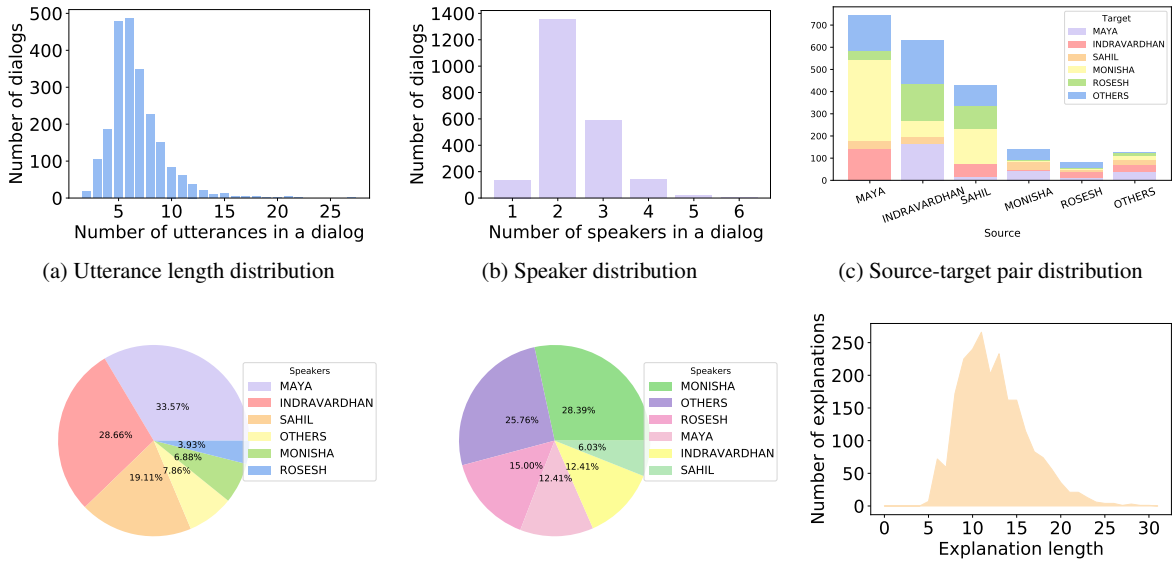
# Dlgs	# Utts	# Eng utts	# Hin utts
2240	9080	101	1453
# CM utts	**Avg. utt/dlg**	**Avg. sp/dlg**	**Avg. words/utt**
7526	4.05	2.35	14.39
Avg. words/dlg	**Vocab size**	**Eng vocab size**	**Hin vocab size**
58.33	10380	2477	7903

Table 1: Statistics of dialogs present in WITS.

3 Dataset

Situational comedies, or *'Sitcoms'*, vividly depict human behaviour and mannerism in everyday real-life settings. Consequently, the NLP research community has successfully used such data for sarcasm identification (Castro et al., 2019; Bedi et al., 2021). However, as there is no current dataset tailored for the proposed task, we curate a new dataset named WITS, where we augment the already existing MASAC dataset (Bedi et al., 2021) with explanations for our task. MASAC is a multimodal, multi-party, code-mixed dialogue dataset compiled from the popular Indian TV show, 'Sarabhai v/s Sarabhai'[2]. We manually analyzed the data and cleaned it for our task. While, the original dataset contained 45 episodes of the TV series, we add 10 more episodes along with their transcription and audio-visual boundaries. Subsequently, we select the sarcastic utterances from this augmented dataset and manually define the utterances to be included in dialogue context for each of them. Finally, we are left with 2240 sarcastic utterances with the number of contextual utterances ranging from 2 to 27. Each of these instances are manually annotated with a corresponding natural language explanation interpreting its sarcasm. Each explana-

[2] https://www.imdb.com/title/tt1518542/



(a) Utterance length distribution (b) Speaker distribution (c) Source-target pair distribution

(d) Sarcasm source distribution (e) Sarcasm target distribution (f) Explanation length distribution

Figure 2: Distribution of attributes in WITS. The number of utterances in a dialog lies between 2 and 27. Maximum number of speakers in a dialogue are 6. The speaker 'Maya' is the most common common sarcasm source while the speaker 'Monisha' is the most prominent sarcasm target.

tion contains four primary attributes – source and target of sarcasm, action word for sarcasm, and an optional description for the satire as illustrated in Figure 1. In the explanation "Indu implies that Maya is not looking good.", 'Indu' is the sarcasm source, 'Maya' is the target, 'implies' is the target word, while 'is not looking good' forms the description part of the explanation. We collect explanations in code-mixed format to keep consistency with the dialogue language. We split the data into train/val/test sets in a 80:10:10 ratio for our experiments which results in 1792 dialogues in the train set and 224 dialogues each in the validation and test sets. More information about the annotation process is present in Appendix A.1. Table 1 and Figure 2 show detailed statistics of WITS.

4 Proposed Methodology

In this section, we present our model and its nuances. The primary goal is to smoothly integrate multimodal knowledge into GPLMs architecture. To this end, we introduce MAF, an adapter-based module that comprises of *Multimodal Context-Aware Attention (MCA2)* and *Global Information Fusion (GIF)* mechanisms. Given the textual input sarcastic instance along with audio-video cues, the former aptly introduces multimodal information in the textual representations, while the latter conglomerates the audio-visual information infused textual representations. This adapter module can be readily incorporated at multiple layers of BART/mBART to facilitate various levels of multimodal interaction. Figure 3 illustrates our model architecture.

4.1 Multimodal Context Aware Attention

The traditional dot-product-based cross-modal attention scheme leads to the direct interaction of textual representations with other modalities. Here the text representations act as the query against the multimodal representations, which serve as the key and value. As each modality comes from a different embedding subspace, a direct fusion of multimodal information might not retain maximum contextual information and can also leak substantial noise in the final representations. Thus, based on the findings of Yang et al. (2019), we propose multimodal fusion through *Context Aware Attention*. We first generate multimodal information conditioned key and value vectors and then perform the traditional scaled dot-product attention. We elaborate on the process below.

Given the intermediate representation H generated by the GPLMs at a specific layer, we calculate the query, key, and value vectors Q, K, and $V \in \mathbb{R}^{n \times d}$, respectively as given in Equation 1, where W_Q, W_K, and $W_V \in \mathbb{R}^{d \times d}$ are learnable parameters. Here, n denotes the maximum sequence length of the text, and d denotes the dimensionality

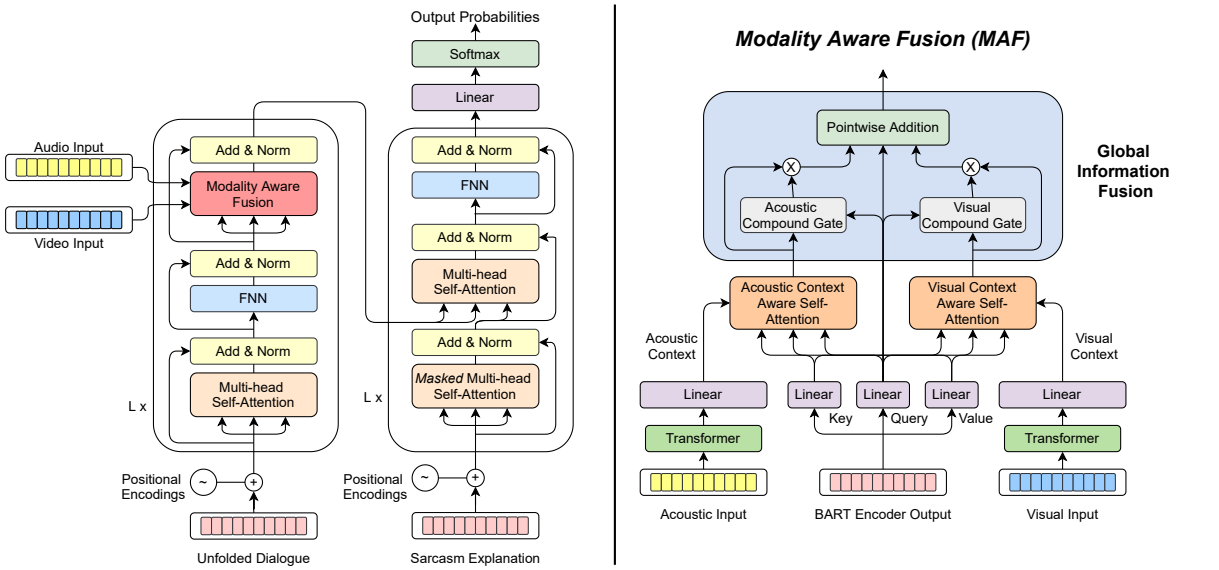


Figure 3: Model architecture for MAF-TAV$_{\text{B}}$. The proposed Multimodal Fusion Block (MFA) captures audio-visual cues using Multimodal Context Aware Attention (MCA2) which are further fused with textual representations using Global Information Fusion (GIF) block.

of the GPLM generated vector.

$$[QKV] = H\,[W_Q W_K W_V] \tag{1}$$

Let $C \in \mathbb{R}^{n_c \times d_c}$ denote the vector obtained from audio or visual representation. We generate multimodal information informed key and value vectors $\hat{Q}$ and $\hat{V}$, respectively, as given by Yang et al. (2019). To decide how much information to integrate from the multimodal source and how much information to retain from the textual modality, we learn $\lambda \in \mathbb{R}^{n \times 1}$ (Equation 3). Note that U_k and $U_v \in \mathbb{R}^{d_c \times d}$ are learnable matrices.

$$\begin{bmatrix}\hat{K}\\ \hat{V}\end{bmatrix} = (1 - \begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix}) \begin{bmatrix}K\\ V\end{bmatrix} + \begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix} (C \begin{bmatrix}U_k\\ U_v\end{bmatrix}) \tag{2}$$

Instead of making λ_k and λ_v as hyperparameters, we let the model decide their values using a gating mechanism as computed in Equation 3. The matrices of $W_{k_1}, W_{k_2}, W_{v_1}$, and $W_{v_2} \in \mathbb{R}^{d \times 1}$ are trained along with the model.

$$\begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix} = \sigma(\begin{bmatrix}K\\ V\end{bmatrix} \begin{bmatrix}W_{k_1}\\ W_{v_1}\end{bmatrix} + C \begin{bmatrix}U_k\\ U_v\end{bmatrix} \begin{bmatrix}W_{k_2}\\ W_{v_2}\end{bmatrix}) \tag{3}$$

Finally, the multimodal information infused vectors $\hat{K}$ and $\hat{V}$ are used to compute the traditional scaled dot-product attention. For our case, we have two modalities – audio and video. Using the *context-aware attention mechanism*, we obtain the acoustic-information-infused and visual-information infused vectors H_A and H_V, respectively (c.f. Equations 4 and 5).

$$H_a = Softmax(\frac{Q\hat{K}_a^T}{\sqrt{d_k}})\hat{V}_a \tag{4}$$

$$H_v = Softmax(\frac{Q\hat{K}_v^T}{\sqrt{d_k}})\hat{V}_v \tag{5}$$

4.2 Global Information Fusion

In order to combine the information from both the acoustic and visual modalities, we design the GIF block. We propose two gates, namely the *acoustic gate* (g_a) and the *visual gate* (g_v) to control the amount of information transmitted by each modality. They are as follows:

$$g_a = [H \oplus H_a]W_a + b_a \tag{6}$$

$$g_v = [H \oplus H_v]W_v + b_v \tag{7}$$

Here, $W_a, W_v \in \mathbb{R}^{2d \times d}$ and $b_a, b_v \in \mathbb{R}^{d \times 1}$ are trainable parameters, and $\oplus$ denotes concatenation. The final multimodal information fused representation $\hat{H}$ is given by Equation 8.

$$\hat{H} = H + g_a \odot H_a + g_v \odot H_v \tag{8}$$

This vector $\hat{H}$ is inserted back into GPLM for further processing.

5 Experiments, Results and Analysis

We illustrate our feature extraction strategy, the comparative systems, followed by the results and its analysis. For a quantitative analysis of the generated explanations, we use the standard metrics

for generative tasks – ROUGE-1/2/L (Lin, 2004), BLEU-1/2/3/4 (Papineni et al., 2002), and METEOR (Denkowski and Lavie, 2014). To capture the semantic similarity, we use the multilingual version of the BERTScore (Zhang et al., 2019).

5.1 Feature Extraction

Audio: Acoustic representations for each instance are obtained using the openSMILE python library[3]. We use a window size of 25 ms and a window shift of 10 ms to get the non-overlapping frames. Further, we employ the eGeMAPS model (Eyben et al., 2016) and extract 154 dimensional functional features such as Mel Frequency Cepstral Coefficients (MFCCs) and loudness for each frame of the instance. These features are then fed to a Transformer encoder for further processing.

Video: We use a pre-trained action recognition model, ResNext-101 (Hara et al., 2018), trained on the Kinetics dataset (Kay et al., 2017) which can recognise 101 different actions. We use a frame rate of 1.5, a resolution of 720 pixels, and a window length of 16 to extract the 2048 dimensional visual features. Similar to audio feature extraction, we employ a Transformer encoder to capture the sequential dialogue context in the representations.

5.2 Comparative Systems

To get the best textual representations for the dialogues, we experiment with various sequence-to-sequence (seq2seq) architectures. **RNN:** We use the openNMT[4] implementation of the RNN seq-to-seq architecture. **Transformer** (Vaswani et al., 2017): The standard Transformer encoder and decoder are used to generate explanations in this case. **Pointer Generator Network** (See et al., 2017): A seq-to-seq architecture that allows the generation of new words as well as copying words from the input text for generating accurate summaries. **BART** (Lewis et al., 2020): It is a denoising autoencoder model with standard machine translation architecture with a bidirectional encoder and an auto-regressive left-to-right decoder. We use its base version. **mBART** (Liu et al., 2020): Following the same architecture and objective as BART, mBART is trained on large-scale monolingual corpora in different languages [5].

[3] https://audeering.github.io/opensmile-python/

[4] https://github.com/OpenNMT/OpenNMT-py

[5] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

Mode	Model	R1	R2	RL	B1	B2	B3	B4	M	BS
Textual	**RNN**	29.22	7.85	27.59	22.06	8.22	4.76	2.88	18.45	73.24
	Transformers	29.17	6.35	27.97	17.79	5.63	2.61	0.88	15.65	72.21
	PGN	23.37	4.83	17.46	17.32	6.68	1.58	0.52	23.54	71.90
	mBART	33.66	11.02	31.50	22.92	10.56	6.07	3.39	21.03	73.83
	BART	36.88	11.91	33.49	27.44	12.23	5.96	2.89	26.65	76.03
Multimodality	**MAF-TA$_M$**	39.02	15.90	36.83	31.26	16.94	11.54	7.72	29.05	77.06
	MAF-TV$_M$	39.47	16.78	**37.38**	32.44	17.91	12.02	7.36	29.74	77.47
	MAF-TAV$_M$	38.52	14.13	36.60	30.50	15.20	9.78	5.74	27.42	76.70
	MAF-TA$_B$	38.21	14.53	35.97	30.58	15.36	9.63	5.96	27.71	77.08
	MAF-TV$_B$	37.48	15.38	35.64	30.28	16.89	10.33	6.55	28.24	76.95
	MAF-TAV$_B$	**39.69**	**17.10**	37.37	**33.20**	**18.69**	**12.37**	**8.58**	**30.40**	**77.67**

Table 2: Experimental results. (Abbreviation: R1/2/L: ROUGE1/2/L; B1/2/3/4: BLEU1/2/3/4; M: METEOR; BS: BERT Score; PGN: Pointer Generator Network).

5.3 Results

Text Based: As evident from Table 2, BART performs the best across all the metrics for the textual modality, showing an improvement of almost 2-3% on the METEOR and ROUGE scores when compared with the next best baseline. PGN, RNN, and Transformers demonstrate admissible performance considering that they have been trained from scratch. However, it is surprising to see mBART not performing better than BART as it is trained on multilingual data. We elaborate more on this in Appendix A.2.

Multimodality: Psychological and linguistic literature suggests that there exist distinct paralinguistic cues that aid in comprehending sarcasm and humour (Attardo et al., 2003; Tabacaru and Lemmens, 2014). Thus, we gradually merge auditory and visual modalities using MAF module and obtain MAF-TAV$_B$ and MAF-TAV$_M$ for BART and mBART, respectively. We observe that the inclusion of acoustic signals leads to noticeable gains of 2-3% across the ROUGE, BLEU, and METEOR scores. The rise in BERTScore also suggests that the multimodal variant is generating a tad bit more coherent explanations. As ironical intonations such as mimicry, monotone, flat contour, extremes of pitch, long pauses, and exaggerated pitch (Rockwell, 2007) form a significant component in sarcasm understanding, we conjecture that our model, to some extent, is able to spot such markers and identify the intended sarcasm behind them.

We notice that visual information also contributes to our cause. Significant performance gains are observed for MAF-TV$_B$ and MAF-TV$_M$, as all the metrics show a rise of about 3-4%. While MAF-TA$_B$ gives marginally better performance over MAF-TV$_B$ in terms of R1, RL, and B1, we see that MAF-TV$_B$ performs better in terms of the rest of the metrics. Often, sarcasm is depicted through gestural cues such as raised eyebrows, a straight face, or an eye roll (Attardo et al.,

Model	R1	R2	RL	B1	B2	B3	B4	M	BS
MAF-TAV$_M$	**39.69**	**17.10**	**37.37**	**33.20**	**18.69**	**12.37**	**8.58**	**30.40**	**77.67**
- MCA2 + CONCAT1	37.56	**14.85**	34.90	30.16	15.76	**10.12**	**6.82**	**28.59**	76.59
- MAF + CONCAT2	17.22	1.70	14.12	13.11	2.11	0.00	0.00	9.34	66.64
- MCA2 + DPA	36.43	13.04	33.75	28.73	14.02	8.00	4.89	25.60	75.58
- GIF	36.37	13.85	34.92	28.49	14.34	9.00	6.16	25.75	**76.86**
MAF-TAV$_B$	**39.69**	**17.10**	**37.37**	**33.20**	**18.69**	**12.37**	**8.58**	**30.40**	**77.67**
- MCA2 + CONCAT1	36.88	13.21	34.39	29.63	14.56	8.43	4.84	26.15	76.08
- MAF + CONCAT2	21.11	2.31	19.68	12.44	2.44	0.73	0.31	9.51	69.54
- MCA2 + DPA	38.84	14.76	36.96	30.23	15.95	9.88	5.83	28.04	77.20
- GIF	39.45	14.85	37.18	31.85	15.97	9.62	5.47	28.87	77.54

Table 3: Ablation results on MAF-TAV$_M$ and MAF-TAV$_B$ (DPA: Dot Product Attention).

2003). Moreover, when satire is conveyed by mocking someone's looks or physical appearances, it becomes essential to incorporate information expressed through visual media. Thus, we can say that, to some extent, our model is able to capture these nuances of non-verbal cues and use them well to normalize the sarcasm in a dialogue. In summary, we conjecture that whether independent or together, audio-visual signals bring essential information to the table to understand sarcasm.

5.4 Ablation Study

Table 3 reports the ablation study. CONCAT1 represents the case where we perform bimodal concatenation $((T \oplus A), (T \oplus V))$ instead of the *MCA2* mechanism, followed by the GIF module, whereas, CONCAT2 represents the simple trimodal concatenation $(T \oplus A \oplus V)$ of acoustic, visual, and textual representations followed by a linear layer for dimensionality reduction. In comparison with *MCA2*, CONCAT2 reports a below-average performance with a significant drop of more than 40% for MAF-TAV$_B$ and MAF-TAV$_M$. This highlights the need to have deftly crafted multimodal fusion mechanisms. CONCAT1, on the other hand, gives good performance and is competitive with DPA and MAF-TAV$_B$. We speculate that treating the audio and video modalities separately and then merging them to retain the complimentary and differential features lead to this performance gain. Our proposed MAF outperforms DPA with gains of 1-3$\%$. This highlights that our unique multimodal fusion strategy is aptly able to capture the contextual information provided by the audio and video signals. Replacing the *GIF* module with simple addition, we observe a noticeable decline in the performance of almost all metrics by about 2-3$\%$. This attests to the inclusion of *GIF* module over simple addition. We also experiment with fusing multimodal information using MAF before different layers of the BART encoder. The best performance was obtained when the fusion was done before the sixth layer of the architecture (c.f. Appendix A.3).

5.5 Result Analysis

We evaluate the generated explanations based on their ability to correctly identify the source and target of a sarcastic comment in a conversation. We report such results for mBART, BART, MAF-TA$_B$, MAF-TV$_B$, and MAF-TAV$_B$. BART performs better than mBART for the source as well as target identification. We observe that the inclusion of audio ($\uparrow 10\%$) and video ($\uparrow 8\%$) information drastically improves the source identification capability of the model. The combination of both these non-verbal cues leads to a whopping improvement of more than 13% for the same. As a result, we infer that multimodal fusion enables the model to incorporate audio-visual peculiarities unique to each speaker, resulting in improved source identification. The performance for target identification, however, drops slightly on the inclusion of multimodality. We encourage future work in this direction.

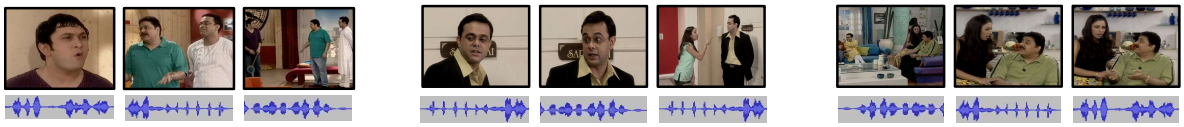
Qualitative Analysis. We analyze the best performing model, MAF-TAV$_B$, and its corresponding unimodal model, BART, and present some examples in Table 4. In Table 4a, we show one instance where the explanations generated by the BART as well as MAF-TAV$_B$ are neither coherent nor comply with the dialogue context and contain much scope of improvement. On the other hand, Table 4b illustrates an instance where the explanation generated by MAF-TAV$_B$ adheres to the topic of the dialogue, unlike the one generated by its unimodal counterpart. Table 4c depicts a dialogue where MAF-TAV$_B$'s explanation better captures the satire than BART. We further dissect the models based on different modalities in Appendix A.4.

Human Evaluation. Since the proposed SED task is a generative task, it is imperative to manually inspect the generated results. Consequently, we perform a human evaluation for a sample of 30 instances from our test set with the help of 25 evaluators[6]. We ask the evaluators to judge the generated explanation, given the transcripts of the sarcastic dialogues along with a small video clip with audio as well. Each evaluator has to see the video clips and then rate the generated explanations on a scale of 0 to 5 based on the following factors[7]:

- **Coherence:** Measures how well the explanations are organized and structured.

[6] Evaluators are the experts in linguistics and NLP and their age ranges in 20-28 years.

[7] 0 denoting poor performance while 5 signifies perfect performance.



	ROSESH: What nonsense? Mujhe kon hunter marega? *What nonsense? Who will beat me with a hunter?* INDRAVARDHAN: Me, me marunga. Kyunki 51 saal baad Maya to mar chuki hogi. Me kisi tarah zinda reh lunga or kahunga ki 'le mar jhadhu, mar, mar jhadhu, mar mar' *I will beat you. Because after 51 years, Maya would be dead. I'll somehow survive and say 'sweep here, sweep here, sweep'*	SAHIL: Ab tumne ghar ki itni saaf safai ki hai and secondly us Karan Verma ke liye pasta, lasagne, caramel custard banaya. *Now you have cleaned the house so much and secondly made pasta, lasagne, caramel custard for that reason Verma.* MONISHA: Walnut brownie bhi. *And walnut brownie too.* SAHIL: Walnut brownie, matlab wo khane wali? *You mean edible walnut brownie?*	MONISHA: Ladki ka naam Ajanta Kyon Rakha? *Why did they named the girl Ajanta?* INDRAVARDHAN: Kyunki uski maa ajanta caves dekh rahi thi Jab vo Paida Hui haha. *Because her mother must be watching the Ajanta caves when she was born haha.*
Gold	Maya Monisha ko tana marti hai safai ka dhyan na rakhne ke liye *Maya taunts Monisha for not keeping a check of cleanliness*	Sahil monisha ki cooking ka mazak udata hai *Sahil makes fun of Monisha's cooking.*	Indravadan Ajanta ke naam ka mazak udata hai *Indravardhan makes fun of Ajanta's name*
BART	Maya Monisha ko tumaari burayi nahi karta. *Maya doesn't blame you for Monisha*	Monisha sahil ko walnut brownie ki matlab wo khane wali. *Walnut Brownie to Monisha Sahil means she eats*	Indravardhan Monisha ko taunt maarta hai ki uski maa ajanta caves dekh rahi thi Jab vo Paida Hui *Indravardhan taunts Monisha as her mother was watching Ajanta Caves when she was born.*
MAF-TAV$_B$	Maya implies ki Monisha bohot ghar mein bahar nahi kar sakati. *Maya implies that Monisha very in home cannot do outside.*	Sahil monisha ki cooking ka mazak udata hai *Sahil makes fun of Monisha's cooking.*	Indravadan ajanta ke naam ka mazak udata hai *Indravardhan makes fun of Ajanta's name*

(a) Incoherant explanation (b) Explanation related to dialogue (c) Explanation related to sarcasm

Table 4: Actual and generated explanations for sample dialogues from test set. The last utterance is the sarcastic utterance for each dialogue.

	mBART	BART	MAF-TA$_B$	MAF-TV$_B$	MAF-TAV$_B$
Source	75.00	77.23	87.94	85.71	**91.07**
Target	45.53	**52.67**	43.75	43.75	46.42

Table 5: Source-target accuracy of the generated explanations for BART-based systems.

- **Related to dialogue:** Measures whether the generated explanation adheres to the topic of the dialogue.
- **Related to sarcasm:** Measures whether the explanation is talking about something related to the sarcasm present in the dialogue.

Table 6 presents the human evaluation analysis with average scores for each of the aforementioned categories. Our scrutiny suggests that MAF-TAV$_B$ generates more syntactically coherent explanations when compared with its textual and bimodal counterparts. Also, MAF-TAV$_B$ and MAF-TV$_B$ generate explanations that are more focused on the conversation's topic, as we see an increase of 0.55 points in the *related to the dialogue* category. Thus, we reestablish that these models are able to incorporate information that is explicitly absent from the dialogue, such as scene description, facial features, and looks of the characters. Furthermore, we establish that MAF-TAV$_B$ is better able to grasp sarcasm and its normalization, as it shows about 0.6 points improvement over BART in the *related to sarcasm* category. Lastly, as none of the metrics in Table 6 exhibit high scores (3.5+), we feel there is still much scope for improvement in terms of the generation performance and human evaluation. The research community can further explore the task with our proposed dataset, WITS.

	Coherency	Related to dialogue	Related to sarcasm
mBART	2.57	2.66	2.15
BART	2.73	2.56	2.18
MAF-TA$_B$	2.95	2.91	2.51
MAF-TV$_B$	3.01	**3.11**	2.66
MAF-TAV$_B$	**3.03**	**3.11**	**2.77**

Table 6: Human evaluation statistics – comparing different models. Multimodal models are BART based.

6 Conclusion

In this work, we proposed the new task of *Sarcasm Explanation in Dialogue* (SED). It aims to generate a natural language explanation for sarcastic conversations. We curated WITS, a novel multimodal, multiparty, code-mixed, dialogue dataset to support the SED task. We experimented with multiple text and multimodal baselines, which give promising results on the task at hand. Furthermore, we designed a unique multimodal fusion scheme to merge the textual, acoustic, and visual features via Multimodal Context-Aware Attention (MCA2) and Global Information Fusion (GIF) mechanisms. As hypothesized, the results show that acoustic and visual features support our task, and thus providing us with better explanations. We show extensive qualitative analysis of the explanations obtained from different models and highlight their advantages as well as pitfalls. We also performed a thorough human evaluation to compare the performance of the models with that of human understanding. Though the models substituted with the proposed fusion strategy perform better than the rest, the human evaluation suggested there is still room for improvement which can be further explored in future studies.

References

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Salvatore Attardo, Jodi Eisterhold, Jennifery Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research*, 16(2).

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, pages 1–1.

Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2017. Harnessing online news for sarcasm detection in hindi tweets. In *Pattern Recognition and Machine Intelligence*, pages 679–686, Cham. Springer International Publishing.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, 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, Online. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.

Herbert L. Colston. 1997. Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes*, 23(1):25–45.

Herbert L Colston and Shauna B Keller. 1998. You'll never believe this: Irony and hyperbole in expressing surprise. *Journal of psycholinguistic research*, 27(4):499–513.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Abhijeet Dubey, Aditya Joshi, and Pushpak Bhattacharyya. 2019. Deep models for converting sarcastic utterances into their non sarcastic interpretation. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, page 289–292, New York, NY, USA. Association for Computing Machinery.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Debanjan Ghosh, Alexander R. Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany. Association for Computational Linguistics.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12972–12980.

Stacey L. Ivanko and Penny M. Pexman. 2003. Context incongruity and irony processing. *Discourse Processes*, 35(3):241–279.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

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, Beijing, China. Association for Computational Linguistics.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6144–6154, Hong Kong, China. Association for Computational Linguistics.

Henri Olkoniemi, Henri Ranta, and Johanna K Kaakinen. 2016. Individual differences in the processing of written sarcasm and metaphor: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(3):433.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319, Saarbrücken, Germany. Association for Computational Linguistics.

Reynier Ortega-Bueno, Francisco Rangel, D Hernández Farıas, Paolo Rosso, Manuel Montes-y Gómez, and José E Medina Pagola. 2019. Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org*, volume 2421, pages 229–256.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Patricia Rockwell. 2007. Vocal features of conversational sarcasm: A comparison of methods. *Journal of psycholinguistic research*, 36(5):361–369.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A corpus of english-hindi code-mixed tweets for sarcasm detection. *arXiv preprint arXiv:1805.11869*.

Sabina Tabacaru and Maarten Lemmens. 2014. Raised eyebrows as gestural triggers in humour: The case of sarcasm and hyper-understanding. *The European Journal of Humour Research*, 2(2):11–31.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm — a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):162–169.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Henry M Wellman. 2014. *Making minds: How theory of mind develops*. Oxford University Press.

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*, WWW '19, page 2115–2124, New York, NY, USA. Association for Computing Machinery.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online. Association for Computational Linguistics.

Baosong Yang, Jian Li, Derek F. Wong, Lidia S. Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):387–394.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Annotation Guidelines

As mentioned in Section 3, we use the MASAC dataset (Bedi et al., 2021) which contains episodes from the series- "Sarabhai v/s Sarabhai" and extract all sarcastic instances from it. Each of these instances is associated with a corresponding video, audio and textual transcript such that the last utterance is sarcastic in nature. We first manually define the number of contextual utterances required to understand the sarcasm present in the last utterance of each instance. Further, we provide each of these sarcastic statements, along with their context, to the annotators which are asked to generate an explanation for these instances based on the audio, video, and text cues. Two annotators were asked to annotate the entire dataset. The target explanation is selected by calculating the cosine similarity between the two explanations. If the cosine similarity is greater than 90% then the shorter length explanation is selected as the target explanation. Otherwise a third annotator goes through the dialogue along the explanations and resolve the conflict. The average cosine similarity after the first pass is 87.67%. All the final selected explanations contain the following attributes:

- Sarcasm source: The speaker in the dialog who is being sarcastic.
- Sarcasm target: The person/thing towards whom the sarcasm is directed towards.
- Action word: Word used to describe how the sarcasm is taking place. For e.g. mock, jokes, etc.
- Description: An optional description about the scene which helps in understanding the sarcasm better.

An example annotation with its attributes is shown in Figure 4.

A.2 Embedding Space for BART and mBART

We compared various text based unimodal methods for our task. Although BART is performing the best for SED, it is important to note that BART is pre-trained on English datasets (GLUE (Wang et al., 2018) and SQUAD (Rajpurkar et al., 2016)). In order to explore how the representation learning is being transferred to a code-mixed setting, we analyse the embedding space learnt by the model before and after fine-tuning it for our task. We considered three random utterances from WITS

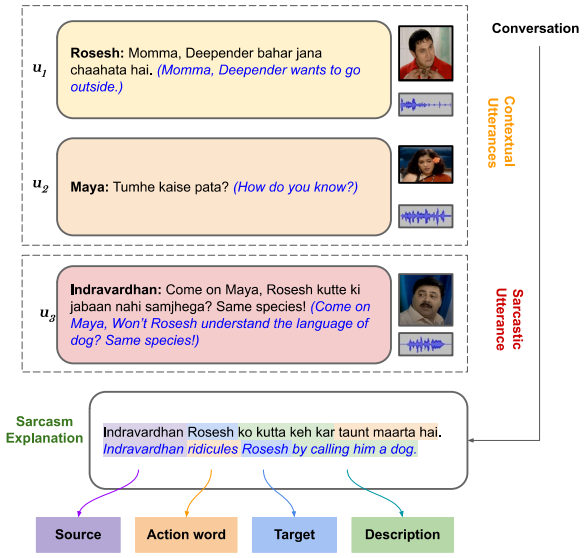


Figure 4: Example annotations from WITS highlighting the different attributes of the explanation.

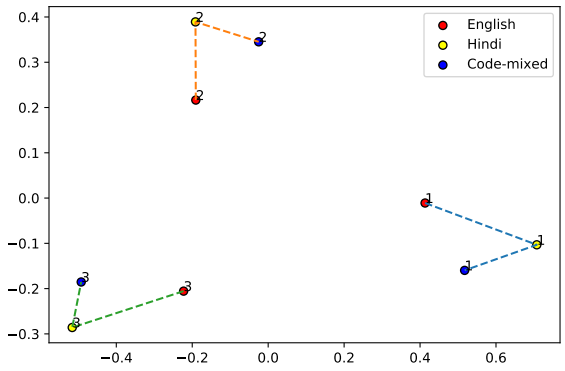
and created three copies of them- one in English, one in Hindi (romanised), and one without modification i.e. code-mixed. Figure 5 illustrates the PCA plot for the embeddings obtained for these nine utterance representations obtained by BART before and after fine-tuning on our task. It is interesting to note that even before any fine-tuning the Hindi, English, and code-mixed representations lie closer to each other and they shift further closer when we fine-tune our model. This phenomenon can be justified as out input is of romanised code-mixed format and thus we can assume that representations are already being captured by the pre-trained model. Fine-tuning helps us understand the Hindi part of the input. Table 7 shows the cosine distance between the representations obtained for English-Hindi, English-Code mixed, and Code mixed-Hindi utterances for the sample utterances. It can be clearly seen that the distance is decreasing after fine-tuning.

Example	English-Hindi		English-Code mixed		Code mixed-Hindi	
	PT	FT	PT	FT	PT	FT
1	0.183	**0.067**	0.014	**0.006**	0.118	**0.056**
2	0.282	**0.093**	0.017	**0.007**	0.197	**0.066**
3	0.321	**0.113**	0.065	**0.020**	0.132	**0.057**

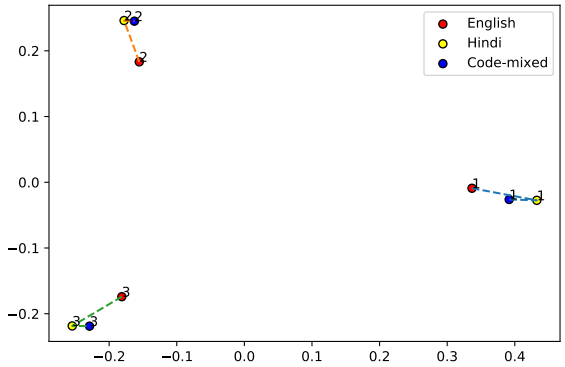
Table 7: Cosine distance between three random samples from the dataset before and after fine-tuning. (PT: pre-trained; FT: fine-tuned)

A.3 Fusion at Different Layers

We fuse the multimodal information of audio and video in the BART encoder using the proposed



(a) Pre-trained



(b) Fine-tuned

Figure 5: Embedding space for BART before and after fine-tuning on sarcasm explanation in dialogs.

fusion mechanism before different layers of the BART encoder. Table 8 shows the results we obtain when the fusion happens at different layers. We obtain the best results when the fusion happens before layer 6 i.e. the last layer of the encoder. This can be attributed to the fact that since there is only one layer of encoder after the fusion, the multimodal information is being retained efficiently and thus being decoded more accurately.

Fusion before layer #	R1	R2	RL
1	37.27	13.95	35.24
2	37.63	14.32	35.57
3	36.73	13.15	34.63
4	37.61	14.98	36.04
5	37.34	13.67	35.48
6	**39.69**	**17.10**	**37.37**

Table 8: ROUGE scores for fusion before different layers (R1/2/L: ROUGE1/2/L).

A.4 More Qualitative Analysis

Table 9a highlights one of many cases where BART is able to capture the essence of sarcasm in a better way when compared to mBART. While mBART gives us an incorrect and incoherent explanation, BART generates an explanation which essentially means the same as the ground truth explanation. The inclusion of audio modality in the unimodal system often helps in generating preferable explanations, as shown in Table 9b. AVII-TA is able to capture the essense of sarcasm in the dialog while the unimodal systems were not able to do so. Furthermore, video modality facilitates even better understanding of sarcasm as illustrated in Table 9c. AVII-TV is able to generate the best results while audio may act as noise in this particular example.

MAYA: Sahil, beta tum bhi soche ho ki maine Monisha ki speech churai? *Sahil, do you also think that I stole Monisha's speech?*
INDRAVARDHAN: Haan.*Yes.*
MAYA: Are darling maine to speech ko chua bhi nahin. chhoti to germs nahin lag jaate? Kyunki Monisha ne mithaai box ki wrapper per likhi thi apni speech hath mein uthati to makkhiya bhanbhana ne lagti. *Darling, I didn't even touch the speech. Would I not have got germs by touching it? Monisha used sweets wrapper to write her speech, if I would have picked it up, there would've been flies buzzing around me.*

Gold	Maya ne Monisha ke speech ka mazak udaya. *Maya makes fun of Monisha's speech.*
mBART	Maya kehti hai ki Monisha ka mazak udata hai *Maya says that make fun of Monisha.*
BART	Maya monisha ke speech ka mazak udati hai *Maya makes fun of Monisha's speech.*
MAF-TA$_B$	Maya monisha ke speech ka mazaak udati hai *Maya says that make fun of Monisha.*
MAF-TV$_B$	Maya mocks monisha kyunki wo rhe theek hai *Maya mocks Monisha because she is okay.*
MAF-TAV$_B$	Maya kehti hai ki uske speech bure hai *Maya says that she didn't like the speech.*

(a) BART v/s mBART: An example where explanation generated by BART is better than mBART.

SAHIL: Ek minute, kya hai maa ji, humaare naatak mein ek bhi stree patra nahi hai, sare ladke hai. *One minute, what is it ma'am, we don't have any female parts in our play, all are male*
PRABHAVATI: To uss mein bhi kaunsi badi baat hai, mai ladka ban jaungi. Mere paas pant shirt to hai, moonche aapki de dena! *So what is the big deal in it, I'll play a male. I have pant shirt, you give me your mustache.*
INDRAVARDHAN: Cancel! Naatak cancel! Maa ji huaa aisa ki humaari jo bahu hai, uska ek chota sa accident ho gaya, to iss liye natak cancel! Monisha le jaao inhe. *Cancel! Play cancel! Ma'am, what happened is, that our daughter in law had a small accident, that is why the play is cancelled. Monisha take her.*
SAHIL: Aur aate aate apna ek chota sa accident bhi kara ke aao! *And when you come, have a small accident too!*

Gold	Sahil Monisha pe gussa hai as usne Prabhavati as an actress le aya. *Sahil is angry on Monisha that she hired Prabhavati as an actress.*
mBART	Sahil ko Prabhavati ko role offer karne par taunt maarta hai. *Sahil taunts because the role is being offered to Prabhavati.*
BART	Indravardhan Monisha ko taunt maarta hai ki uska ek chota sa accident bhi kara ke aao. *Indravardhan taunts Monisha that she should have a small accident.*
MAF-TA$_B$	Sahil ko Prabhavati ko role offer nahi karna. *Sahil does not want Prabhavati to have this role.*
MAF-TV$_B$	Sahil Indravardhan ko ek accident keh ke uska majaak udaata hai. *Calls Indravardhan an accident and makes fun of him.*
MAF-TAV$_B$	Sahil ko Prabhavati ko role offer nahi karna. *Sahil does not want Prabhavati to have this role.*

(b) Audio helps: An example where audio modality helps in generating more fitting explanation.

MAYA: Kshama? You mean Sahil Kshama ko pyaar karta hai!? *Kshama? You mean Sahil loves Kshama?*
SAHIL: Nahi, nahi! Ek minute, ek minute, mai kshama chahata hu. *No no, One minute, one minute, I want forgiveness (kshama in hindi).*
INDRAVARDHAN: Dekha, Kshama chahata hai! Chahata ka matlab pyaar karna hi hua na!? *See, wants forgiveness! Wants means love only, no!?*

Gold	Indravardhan Sahil ko tease karta hai ki vo Kshama se pyaar karta hai.*Indravardhan teases Sahil by implying that he loves kshama (name of a girl in hindi meaning forgiveness)*
mBART	Indravardhan implies ki Sahil ek kshama chahata hai. *Indravardhan implies that Sahil wants forgiveness.*
BART	Maya ko kshama chahata hai *Maya wants forgiveness.*
MAF-TA$_B$	Indravardhan Kshama ko pyaar karne par taunt maarta hai. *Indravardhan taunts that he loves Kshama.*
MAF-TV$_B$	Indravardhan majaak mein kehta hai ki Sahil Kshama ko pyaar karta hai. *Indravardhan jokes that Sahil loves Kshama*
MAF-TAV$_B$	Indravardhan Rosesh ko Kshama ki matlab pyaar karne par taunt maarta hai. *Indravardhan taunts Rosesh for loving the meaning of forgiveness.*

(c) Video helps: An example where video modality helps in generating more fitting explanation.

MAYA: And this time I thought lets have a theme party! animals! Hum log sab animals banenge! *And this time I thought lets have a theme party! animals! We will all be animals!*
MONISHA: Walnut brownie bhi. *And walnut brownie too.*
MAYA: Mai hiran, Sahil horse, and Monisha chhipakalee! *I'll be a deer, Sahil horse, and Monisha lizard!*

Gold	Maya Monisha ko chhipakalee keha kar uska majaak udaati hai.*Maya makes fun of Monisha by comparing her with a lizard.*
mBART	Maya Monisha ko taunt maarti hai ki use animal themed party *Maya taunts Monisha for her animal themed party.*
BART	Maya Monisha ko taunt maarti hai. *Maya taunts Monisha.*
MAF-TA$_B$	Maya implies ki vo animal mein theme party ke baare mein nahi banenge. *Maya implies that she won't be in regarding animal themed party.*
MAF-TV$_B$	Maya Monisha ke animal ke behaviour par taunt maarti hai. *Maya taunts Monisha for her animal behaviour.*
MAF-TAV$_B$	Maya Monisha ko animal kaha ke taunt maarti hai. *Maya taunts Monisha by calling her an animal.*

(d) Audio and video helps: An example where audio and video modality together helps in generating better explanation.

Table 9: Actual and generated explanations for sample dialogues from test set. The last utterance is the sarcastic utterance for each dialogue.