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), 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, barring 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\(_B\) and MAF-TAV\(_M\) 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
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 conversation 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.

### 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 explanation

---

\(^2\)https://www.imdb.com/title/tt1518542/
target 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.](image)

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...
of the GPLM generated vector.

\[
[QKV] = H \left[ W_QW_KW_V \right] \quad (1)
\]

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 \( Q \) and \( V \), 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.

\[
\begin{bmatrix} \hat{K} \\ \hat{V} \end{bmatrix} = (1 - \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} \begin{bmatrix} K \\ V \end{bmatrix}) + \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} \begin{bmatrix} C \left[ U_k \right] \\ C \left[ U_v \right] \end{bmatrix} \quad (2)
\]

Instead of making \( \lambda_k \) and \( \lambda_v \) as hyperparameters, we let the model decide their values using a gating mechanism as computed in Equation 3. The matrices of \( W_{k1}, W_{k2}, W_{v1}, \) and \( W_{v2} \in \mathbb{R}^{d \times d} \) are trained along with the model.

\[
\begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} = \sigma \left( \begin{bmatrix} K \\ V \end{bmatrix} \begin{bmatrix} W_{k1} \\ W_{v1} \end{bmatrix} + C \left[ U_k \right] \begin{bmatrix} W_{k2} \\ W_{v2} \end{bmatrix} \right) \quad (3)
\]

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).

\[
H_A = \text{Softmax} \left( \frac{Q \hat{K}_A^T}{\sqrt{d_k}} \right) \hat{V}_A \quad (4)
\]

\[
H_V = \text{Softmax} \left( \frac{Q \hat{K}_V^T}{\sqrt{d_k}} \right) \hat{V}_V \quad (5)
\]

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:

\[
g_a = [H \oplus H_A]W_a + b_a \quad (6)
\]

\[
g_v = [H \oplus H_V]W_v + b_v \quad (7)
\]

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.

\[
\hat{H} = H + g_a \odot H_A + g_v \odot H_V \quad (8)
\]

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 library⁷. 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, ResNet-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 openNMT⁸ 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 languages⁹.

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 and MAF-TAVₘ 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-TVₐ and MAF-TVₐₘ as all the metrics show a rise of about 3-4%. While MAF-TAₐ gives marginally better performance over MAF-TVₐ in terms of R1, RL, and B1, we see that MAF-TVₐₘ 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.,...
we observe a noticeable decline in the performance with a significant drop of more than we conjecture that whether independent or together, this highlights that our unique multimodal fusion mechanisms. C to have deftly crafted multimodal fusion mechanisms. CONCAT1, on the other hand, gives good performance and is competitive with DPA and MAF-TAV. 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).

### Table 3: Ablation results on MAF-TAVₘ and MAF-TAVₐ (DPA: Dot Product Attention).

<table>
<thead>
<tr>
<th>Model</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>B5</th>
<th>B6</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF-TAVₘ</td>
<td>28.82</td>
<td>29.00</td>
<td>29.77</td>
<td>29.80</td>
<td>30.54</td>
<td>29.80</td>
</tr>
<tr>
<td>- MCA2 + CONCAT1</td>
<td>28.82</td>
<td>29.00</td>
<td>29.77</td>
<td>29.80</td>
<td>30.54</td>
<td>29.80</td>
</tr>
<tr>
<td>- MAF + CONCAT2</td>
<td>28.82</td>
<td>29.00</td>
<td>29.77</td>
<td>29.80</td>
<td>30.54</td>
<td>29.80</td>
</tr>
<tr>
<td>- MAF + DPA</td>
<td>28.82</td>
<td>29.00</td>
<td>29.77</td>
<td>29.80</td>
<td>30.54</td>
<td>29.80</td>
</tr>
<tr>
<td>- GIF</td>
<td>28.82</td>
<td>29.00</td>
<td>29.77</td>
<td>29.80</td>
<td>30.54</td>
<td>29.80</td>
</tr>
</tbody>
</table>

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-TAVₐ and MAF-TAVₘ. 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. 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-TAVₐ, MAF-TAVₘ, and MAF-TAVₐ. 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-TAVₐ, 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-TAVₐ 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-TAVₐ adheres to the topic of the dialogue, unlike the one generated by its unimodal counterpart. Table 4c depicts a dialogue where MAF-TAVₐ’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. 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:

- **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 Denoting poor performance while 5 signifies perfect performance.
Table 6: Human evaluation statistics — comparing different models. Multimodal models are BART based.

<table>
<thead>
<tr>
<th></th>
<th>mBART</th>
<th>BART</th>
<th>MAF-TAV_8</th>
<th>MAF-TV_8</th>
<th>MAF-TAV_8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>75.00</td>
<td>77.23</td>
<td>87.94</td>
<td>85.71</td>
<td>91.07</td>
</tr>
<tr>
<td>Target</td>
<td>45.53</td>
<td>52.67</td>
<td>43.75</td>
<td>43.75</td>
<td>46.42</td>
</tr>
<tr>
<td>Coherence</td>
<td>2.57</td>
<td>2.66</td>
<td>2.15</td>
<td>2.18</td>
<td>2.51</td>
</tr>
<tr>
<td>Related to dialogue</td>
<td>0.72</td>
<td>0.76</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Related to sarcasm</td>
<td>0.72</td>
<td>0.76</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
</tbody>
</table>

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


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.

![Figure 4: Example annotations from WITS highlighting the different attributes of the explanation.](image)

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 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 our 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.

<table>
<thead>
<tr>
<th>Example</th>
<th>English-Hindi</th>
<th>English-Code mixed</th>
<th>Code mixed-Hindi</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>FT</td>
<td>PT</td>
<td>FT</td>
</tr>
<tr>
<td>1</td>
<td>0.183</td>
<td>0.067</td>
<td>0.014</td>
</tr>
<tr>
<td>2</td>
<td>0.282</td>
<td>0.093</td>
<td>0.017</td>
</tr>
<tr>
<td>3</td>
<td>0.321</td>
<td>0.113</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>0.006</td>
<td>0.020</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>0.118</td>
<td>0.057</td>
<td></td>
</tr>
</tbody>
</table>

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
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 essence 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.

![Figure 5: Embedding space for BART before and after fine-tuning on sarcasm explanation in dialogs.](image)

 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.

<table>
<thead>
<tr>
<th>Fusion before layer #</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37.27</td>
<td>13.95</td>
<td>35.24</td>
</tr>
<tr>
<td>2</td>
<td>37.63</td>
<td>14.32</td>
<td>35.57</td>
</tr>
<tr>
<td>3</td>
<td>36.73</td>
<td>13.15</td>
<td>34.63</td>
</tr>
<tr>
<td>4</td>
<td>37.61</td>
<td>14.98</td>
<td>36.04</td>
</tr>
<tr>
<td>5</td>
<td>37.34</td>
<td>13.67</td>
<td>35.48</td>
</tr>
<tr>
<td>6</td>
<td>39.69</td>
<td>17.10</td>
<td>37.37</td>
</tr>
</tbody>
</table>

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
MAYA: Sahil, beta tum hi sochte ho ki maine Monisha ki speech chunai? Sahil, do you also think that I stole Monisha’s speech? 

INDRAVARDHAN: Haan Yes.

MAYA: Are darling maine to speech ko chua hai bhi nahiin. chhoti chhoti speech ko maazak udati hai. Maya mocks Monisha because she is okay. 

Maya: One minute, one minute, I want forgiveness (kshama in hindi).

SAHIL: Nahi, nahi! Ek minute, ek minute, mai kshama chahata hu. Indravardhan taunts Sahil by implying that he loves kshama (name of a girl in hindi meaning forgiveness). 

MAYA: Kshama? You mean Sahil Kshama ko pyaar karta hai!?

MAYA: Are darling maine to speech ko chua bhi nahin. chhoti chhoti speech ko maazak udati hai. Maya mocks Monisha because she is okay.

MAYA: No no, One minute, one minute, I want forgiveness (kshama in hindi).

SAHIL: Nahi, nahi! Ek minute, ek minute, mai kshama chahata hu. Indravardhan taunts Sahil by implying that he loves kshama (name of a girl in hindi meaning forgiveness). 

MAYA: Are darling maine to speech ko chua bhi nahin. chhoti chhoti speech ko maazak udati hai. Maya mocks Monisha because she is okay.

SAHIL: Ek minute, kya hai maa ji, humaare naatak mein ek bhi stree patra nahi hai, sare laikhe 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 bad baat hai, mai laikda ban jaungi. Mere paas pant shirt to hai, moonee 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 bhaa aisa ki humaati jo bahtu hai, uska ek chota sa accident ho gaya, to iss liye nataak cancel! Monisha le jaao inhe. Cancel! Play cancel! Maa’um, 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?

MAYA: Are darling maine to speech ko chua bhi nahin. chhoti chhoti speech ko maazak udati hai. Maya mocks Monisha because she is okay.

Prabhavati: To uss mein bhi kaunsi bad baat hai, mai laikda ban jaungi. Mere paas pant shirt to hai, moonee 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 bhaa aisa ki humaati jo bahtu hai, uska ek chota sa accident ho gaya, to iss liye nataak cancel! Monisha le jaao inhe. Cancel! Play cancel! Maa’um, 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?

(b) Audio helps: An example where audio 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! 

SAHIL: Ek minute, kya hai maa ji, humaare naatak mein ek bhi stree patra nahi hai, sare laikhe 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 bad baat hai, mai laikda ban jaungi. Mere paas pant shirt to hai, moonee 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.

MAYA: Are darling maine to speech ko chua bhi nahin. chhoti chhoti speech ko maazak udati hai. Maya mocks Monisha because she is okay.

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! 

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

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