Well, Now We Know! Unveiling Sarcasm: Initiating and Exploring Multimodal Conversations with Reasoning

Gopendra Vikram Singh\textsuperscript{1,*}, Maujama Firdaus\textsuperscript{2,*}, Dushyant Singh Chauhan\textsuperscript{1,*}, Asif Ekbal\textsuperscript{1}, Pushpak Bhattacharyya\textsuperscript{3}

\textsuperscript{1}Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
\textsuperscript{2}Department of Computing Science, University of Alberta, Canada
\textsuperscript{3}Indian Institute of Technology Bombay, India

\{gopendra.99, mauzama.03, dushyantchauhan27, pushpakbh\}@gmail.com, asif@iitp.ac.in

Abstract

Sarcasm is a widespread linguistic phenomenon that poses a considerable challenge to explain due to its subjective nature, absence of contextual cues and rooted personal perspectives. Even though the identification of sarcasm has been extensively studied in dialogue analysis, merely detecting sarcasm falls short of enabling conversational systems to genuinely comprehend the underlying meaning of a conversation and generate fitting responses. It is imperative to not only detect sarcasm but also pinpoint its origination and the rationale behind the sarcastic expressions to capture its authentic essence. In this paper, we delve into the discourse structure of conversations infused with sarcasm and introduce a novel task - Sarcasm Initiation and Reasoning in Conversations (SIRC). Embedded in a multimodal environment and involving a combination of both English and code-mixed interactions, the objective of the task is to discern the trigger or starting point of sarcasm. Additionally, the task involves producing a natural language explanation that rationalizes the satirical dialogues. To achieve this, we introduce Sarcasm Initiation and Reasoning Dataset (SIRD) to facilitate our task and provide sarcasm initiation annotations and reasoning. We develop a comprehensive model named Sarcasm Initiation and Reasoning Generation (SIRG), which is designed to encompass textual, audio, and visual representations. To achieve this, we introduce a unique shared fusion method that employs cross-attention mechanisms to seamlessly integrate these diverse modalities. Our experimental outcomes, conducted on the SIRD dataset, demonstrate that our proposed framework establishes a new benchmark for both sarcasm initiation and its reasoning generation in the context of multimodal conversations. The code and dataset can be accessed from \url{https://www.iitp.ac.in/~ai-nlp-ml/resources.html#sarcasm-explain} and \url{https://github.com/GussailRaat/SIRG-Sarcasm-Initiation-and-Reasoning-Generation}.

Introduction

Sarcasm is an enduring linguistic phenomenon that presents a significant challenge to elucidate, owing to its subjective nature, absence of contextual cues, and profound underlying sentiments. Sarcasm denotes the utilization of satirical or ironic expressions, often intended to inflict hurt, insult, or provoke offense. The apparent meaning of these statements typically contrasts with their intended meaning. Understanding sarcasm necessitates an awareness of the context in which the statement was articulated.

(Joshi, Sharma, and Bhattacharyya 2015) proposed that the existence of incongruity serves as a crucial indicator of sarcasm. Conventional investigations into sarcasm analysis have primarily focused on identifying latent sarcasm within text (Campbell and Katz 2012). Over the past few years, there has been a growing trend in utilizing multimodal signals, such as images, videos, and audio to detect sarcasm (Schifanella et al. 2016; Castro et al. 2019). By incorporating multimodal signals, the realm of incongruity within sarcastic content broadens to encompass both inter-modality and intra-modality incongruity. Many current systems depend on the interplay of latent representations specific to each modality to harness this incongruity.

While the utilization and comprehension of sarcasm pose cognitive challenges (Olkoniemi, Ranta, and Kaakinen 2016), psychological research suggests a positive link between sarcasm and the receiver’s theory of mind (ToM), signifying the capacity to interpret another individual’s mental state (Wellman 2014). To enable NLP systems to replicate such human-like intelligent behavior, they must not only excel at sarcasm detection but also demonstrate the capability to know the initiation point and fully understand it. With this objective, progressing beyond sarcasm identification, we introduce the innovative task of Sarcasm Initiation and Reasoning in Conversations.

Sarcasm initiation refers to the act of starting or initiating a sarcastic statement, remark, or expression. Sarcasm itself is a form of verbal irony where someone says something but means the opposite, often in a humorous or mocking way. When discussing “sarcasm initiation,” it is about identifying the moment or context when a person begins to communicate using sarcasm. This initiation might involve tone of voice, choice of words, or other cues that indicate the shift from straightforward communication to sarcastic commentary.

In conversations or dialogues, the cause or span of sarcasm can vary. It can be a single sentence, phrase, or even a single word used sarcastically. It may also extend to multi-
The main contributions of our work can be summarized as follows. First, we introduce a novel task named Sarcasm Initiation and Reasoning in Conversations (SIRC) where we intend to extract the spans for the cause of sarcasm and the reasoning behind them in addition to purely identifying the sarcastic utterances. Second, we provide the first multimodal conversational corpus, Sarcasm Initiation and Reasoning Dataset (SIRD), which includes annotations for cause of sarcasm, sarcasm reasoning, speaker information. Third, we propose SIRG, a multimodal multitask system that incorporates shared fusion mechanism for sarcasm detection, its initiation spans and sarcasm reasoning in conversational data. Lastly, experimental results show performance improvement compared to the baselines and provide a benchmark for the said task.

**Related Work**

Historically, the identification of sarcasm has primarily depended on rule-based classification techniques (Joshi, Bhatcharyya, and Carman 2017; Veale and Hao 2010). Nevertheless, a distinct approach was pursued by Poria et al. (2016), who utilized sentiment and emotion features derived from pre-trained models focused on sentiment, emotion, and personality within a textual dataset.

Over the past few years, researchers (Castro et al. 2019) have initiated investigations into harnessing the capabilities of multimodal information sources in the domain of sarcasm detection. Remarkably, the pioneering work by Castro et al. (2019) marked the inception of the MUStARD dataset, uniquely tailored for the purpose of sarcasm detection. Delving further, Kumar et al. (2022) delve into dissecting the discourse structure within sarcastic conversations and introduce the Sarcasm Explanation in Dialogue (SED) task. In a similar vein, Chauhan et al. (2020a) put forth two attention-like mechanisms, the Inter-task Relationship Module (iTRM) and the Inter-class Relationship Module (iCRM), to comprehend the connections and resemblances between the tasks involving sarcasm, emotion, and sentiment. Moreover, Babanejad et al. (2020) introduce two novel deep neural network models, ACE 1 and ACE 2, designed to process a text passage as input and predict its sarcasm presence.

![A figure showing examples of SIRD-English and SIRD-Hinglish datasets showing sarcasm initiation and reasoning](Image)

Figure 1: Example of SIRD-English and SIRD-Hinglish datasets showing sarcasm initiation and reasoning.
from previous endeavors, as we present a cohesive task that uncovers the nuanced aspects of sarcasm. This involves not only pinpointing the starting point of sarcasm but also constructing the rationale behind a specific sarcastic utterance within its multimodal context.

**Methodology**

In this section, we present our Sarcasm Initiation and Reasoning Generation (SIRG) framework and its key features, aiming to seamlessly integrate multimodal knowledge into the BART architecture. We introduce a module called Multimodal Shared Fusion (MSF), which consists of two mechanisms: CoMat and Attention Map. With the inclusion of both textual input containing sarcastic dialogue and audio-video cues, the former mechanism effectively incorporates multimodal information into the textual representations. Meanwhile, the latter mechanism consolidates the audio-visual information into textual representations that have been enhanced with multimodal data. Our adapter module can be easily integrated at various layers of BART/mBART, enabling different levels of multimodal interaction. Please refer to Figure 2 for a visual representation of our model architecture.

**Audio Encoder.** To extract acoustic features, we utilize OpenSMILE (Eyben, Wöllmer, and Schuller 2010). This tool employs a diverse range of filters, capabilities, and transformations to extract Low-Level Descriptors (LLD) and perform various manipulations on them.

**Video Encoder.** The visual world and facial expressions offer rich emotional indicators. After obtaining the visual embedding, we separated the embedding dimensions and dataset into groups to simplify the problem and make better use of the complete embedding space. Each learner will create a unique distance metric using just a subspace of the original embedding space and a portion of the training data. By segmenting the network’s embedding layer into D consecutive slices, we are able to isolate D unique learners inside the embedding space. After learners’ solution converge, we aggregate them to obtain the whole embedding space. The merging is accomplished by recombining the slices of the embedding layer that corresponds to the D learners. To ensure uniformity in the embeddings produced by the various learners, we then perform fine-grained tuning across the entire dataset. The merged embeddings may be hampered by the gradients, which resemble white noise and would hinder training performance. This is called the ”shattered gradients problem”. To address this, residual weights (Balduzzi et al. 2017) provide the gradients with some spatial structure, which aids in training, as shown in Figure 2.

**Shared-Fusion.** In the context of sarcasm cause and its reasoning detection, multimodality plays a vital role. To address this task, we utilize a method called cross-attention-based fusion Shared-Fusion. This method enables effective integration of multiple modalities by capturing inter-modal information while preserving intra-modal features. By combining and attending to features from audio, visual, and textual modalities, we create a unified feature representation. The Shared-Fusion method employs cross-attention to encode the inter-modal relationships among different modalities. This allows the model to capture relevant information from each modality and generate a comprehensive representation. The intra-modal features are preserved to retain the unique characteristics of each modality.

The fusion process involves concatenating the features from audio, visual, and textual modalities and attending to them to capture the important cues for sarcasm and its cause detection. This combined feature representation provides a holistic view of the multimodal input, enabling better understanding and detection of sarcasm and its underlying cause. By leveraging multimodal fusion and cross-attention mechanisms, we aim to seamlessly integrate multimodal knowledge into the BART framework and its key features, offering rich multimodal representations that capture the unique characteristics of each modality.
nisms, the Shared-Fusion method enhances the performance of sarcasm and cause detection by effectively integrating information from multiple modalities.

**Shared-Features.** Let us assume that $F_t$, $F_a$, and $F_v$ correspond to the feature vectors of text, audio, and video. As shown in Figure 3, the text, audio, and video feature vectors are concatenated to provide the representation of text, audio, and video features, $Z = [F_t; F_a; F_v] \in \mathcal{R}^{D\times L}$, where $D$ denoted by $D = x_t + x_a + x_v$ shows the concatenated features (text, audio, video) dimension. Given the multimodal utterance ($M_o$), the combined feature representations ($Z$) are now used to focus on the unimodal feature representations $F_t$, $F_a$, and $F_v$. The combined features ($Z$) and the shared correlation matrix ($M$) between the text features are given by:

$$CoMat = tanh^{FTW_{za}}$$

Where $W_{za} \in \mathcal{R}^{L\times L}$ is the learnable weight matrix across text and shared textual, audio, and video features. Similarly correlation matrix for audio and video features is:

$$CoMat = tanh^{FTW_{va}}$$

$$CoMat = tanh^{FTW_{va}}$$

The shared correlation matrices $CoMat_t$, $CoMat_a$, and $CoMat_v$ for the text, audio, and video modalities offer a semantic indicator of importance both within and between modalities. Within the same modality, there is a high correlation between the matching samples and the other modalities, as indicated by a greater correlation coefficient of the shared correlation matrices $CoMat_t$, $CoMat_a$, and $CoMat_v$. In order to improve the performance of the system, the suggested strategy effectively takes advantage of the complementary nature of the text, audio, and video modalities (i.e., inter-modal relationships) and intra-modal relationships. The attention weights of the audio, text, and video modalities are computed following the computation of the shared correlation matrices.

We use several learnable weight matrices corresponding to features of the separate modalities to compute attention weights for the modalities because the dimensions of shared correlation matrices and the features of the associated modality vary. The learnable weight matrices $W_t$ and $W_t$ are used to combine the shared correlation matrix $CoMat_t$ and the matching textual features $F_t$, the following formula is used to calculate the attention weights for the textual modality:

$$Atte_t = ReLu(W_tF_t + W_tCoMat_t^T)$$

where $Atte_t$ represents the attention map of textuality. Similarly, for visual and acoustic attention map the equations are:

$$Atte_v = ReLu(W_vF_v + W_vCoMat_v^T)$$

$$Atte_a = ReLu(W_aF_a + W_aCoMat_a^T)$$

The attended characteristics of text, audio, and video modalities are computed using the attention maps. These characteristics are attained by:

$$X_{Atte_t} = WAtte_t + F_t$$

$$X_{Atte_a} = WAtte_a + F_a$$

$$X_{Atte_v} = WAtte_v + F_v$$

The shared features of attended features of audio, video, and text is obtained by:

$$X_{Atte} = [X_{Atte_t}; X_{Atte_a}; X_{Atte_v}]$$

**Calculation of Final Loss.** To unify the various losses in our framework, we combine them into a single unified loss function:

$$\mathcal{L} = \mathcal{L}_{SD} + \mathcal{L}_{SI} + \mathcal{L}_{SR}$$

Where SD: Sarcasm Detection, SI: Sarcasm Initiation and SR: Sarcasm reasoning.

**Dataset.**

Situation comedies, commonly known as ‘Sitcoms’, vividly portray human behavior and interactions within everyday real-life contexts. As a result, the NLP research field has effectively leveraged such datasets to achieve sarcasm identification (Castro et al. 2019; Bedi et al. 2021). Given the absence of an existing dataset suitable for our intended task, we compile a new dataset called Sarcasm Initiation and Reasoning Dataset (SIRD) having two sub-datasets in English and in code-mixed (Hindi-English). For creating this dataset, we enhance the pre-existing English MUSTARD dataset (Castro et al. 2019) and code-mixed MASAC dataset (Bedi et al. 2021) by incorporating reasoning and marking the sarcasm initiation spans tailored to our specific task.

**Dataset Description.** The MUSTARD dataset (Castro et al. 2019) encompasses audio-visual utterances derived from dialogues, amounting to a cumulative duration of 3.68 hours. Comprising 690 samples, each instance incorporates an utterance, its corresponding context, and a label denoting its sarcastic or non-sarcastic nature. The dataset was meticulously curated, drawing samples from acclaimed TV series including Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics Anonymous.

MASAC (Bedi et al. 2021) is an amalgamated, multimodal, multi-party dialogue dataset in Hindi-English code-mixed format, sourced from the renowned Indian TV series Sarabhai vs Sarabhai. Starting from the initial dataset of 45 TV series episodes, we expand it by including an additional 15 episodes, complete with their transcriptions and audio-visual divisions. From this extended dataset, we meticulously choose the sarcastic utterances and manually determine the specific dialogues that should surround each of these sarcastic instances. The outcome is a collection of 2255 sarcastic dialogues, with the count of contextual utterances varying between 2 and 25.

**Dataset Annotation.** We undertake a manual analysis of the data and perform necessary cleaning tailored to our task. For sarcasm initiation labeling, every entry within both the dataset comprises a single sentence serving as the utterance,
accompanied by numerous antecedent sentences forming the contextual backdrop of a dialogue. For dataset enhancement, we executed a manual re-annotation process, encompassing both sarcasm labels and the underlying initiation spans of the sarcasm. Our annotation guidelines were founded on the prior work of (Poria et al. 2021; Ghosh et al. 2022).

For ensuring precision, two knowledgeable human experts, both graduate students well-versed in the task, autonomously annotated each utterance. The definitive causal span was determined from the composite of candidate spans proposed by different annotators. However, this was considered only if the overlap between their spans was at least 50% of the size of the smallest candidate span. In instances where a consensus couldn’t be achieved based on prior spans, a third annotator was engaged. This third annotator adhered to a guideline favoring shorter spans as long as they comprehensively conveyed the essence of the sarcasm without any loss of pertinent information.

For sarcasm reasoning, each instance is associated within the SIRD dataset is linked to a corresponding video, audio, and textual transcript, with the final utterance consistently being of sarcastic nature. Initially, we manually establish the count of contextual utterances necessary to comprehend the sarcasm expressed in the concluding utterance of each dialogue. Additionally, we present each of these sarcastic statements, in conjunction with their context, to the annotators, tasking them with producing a reasoning for these instances guided by cues from audio, video, and text sources. Two annotators were assigned to annotate the complete dataset. The desired explanation is chosen by evaluating the cosine similarity between the two explanations. If the cosine similarity surpasses 85%, the fluent explanation is chosen as the target. If not, a third annotator reviews the dialogue alongside the explanations to settle any discrepancies. Following the initial assessment, the average cosine similarity stands at 82.48%. Every chosen reasoning includes the speaker information, the individual towards which the sarcasm is aimed and cause of sarcasm with the justification to explain why the given utterance is sarcastic. Figure 1 depicts an example annotation of both sarcasm initiation span and reasoning from the SIRD dataset along with its associated attributes.

**Experiments**

In this section, we present the implementation details, followed by the baselines for all the tasks and the evaluation metrics for all three tasks.

**Experimental Setup.** We use PyTorch\(^2\), a Python-based deep learning package, to develop our proposed model. We conduct experiments with the BART import from the huggingface transformers\(^3\) package. To establish the ideal value of the additive angle \(x\), which affects performance, five values ranging from 0.1 to 0.5 were examined. The default value for \(x\) is 0.30. We set amplification value \(a\) as 64. All experiments are carried out on an NVIDIA GeForce RTX 2080 Ti GPU. We conducted a grid search across 200 epochs.

\(^{2}\)https://pytorch.org/
\(^{3}\)https://huggingface.co/docs/transformers/index

We find empirically that our Embedding size is 812 bytes. We use Adam (Kingma and Ba 2015) for optimization. The learning rate is 0.05, and the dropout is 0.5. The auto-latent encoder’s dimension is fixed at 812. The discriminator \(D\) consists of two completely linked layers and a ReLU layer and accepts 812-D input features. Stochastic gradient descent has a learning rate of 1e-4 and a weight decay of 1e-3, with a momentum of 0.5.

**Baselines.** We discuss the details of the considered baselines below. Similar to the SIRG approach, to adapt the baselines to our multi-task scenario, we add a linear layer on top of the hidden-states output in the output layer of the CE task to calculate span start and end logits. The output layer for the CE task employs sigmoid activation, in which the threshold value is set at 0.4. For sarcasm initiation we follow the standard baselines such as (i) BiRNN-Attention (Liu and Lane 2016), (ii) CNN-GRU Zhang, Robinson, and Tepper (2018), (iii) BERT (Liu et al. 2019), (iv) BiRNN-HateXplain and BERT-HateXplain Mathew et al. (2021), (v) SpanBERT (Joshi et al. 2020), (vi) Cascaded Multitask System with External Knowledge Infusion (CMSEKI) (Ghosh, Ekbal, and Bhattacharyya 2022). For sarcasm reasoning, we employ encoder-decoder baselines, such as (i) Pointer Generator Network (See, Liu, and Manning 2017), (ii) BART (Lewis et al. 2019), (iii) mBART (Liu et al. 2020), (iv) MAF\(_{TAV}\), MAF\(_{B}\) (Kumar et al. 2022).

**Evaluation Metrics.** To ensure a comprehensive comparison across all tasks, we conduct both automatic and human evaluations. In the case of sarcasm detection, widely accepted metrics such as Accuracy and F1 score are utilized. For the sarcasm initiation task, we employ a range of metrics including F-Measure-Modified (FM), Precision-Modified (PM), Hamming Distance (HD), Jaccard F1 (JF), and Recall-Oriented Score (ROS). For reasoning generation, we resort to standard generative task metrics like ROUGE-L, BLEU-3/4, and METEOR, and incorporate the multilingual version of BERTScore to gauge semantic similarity.

In the context of manual evaluation, we randomly select 250 model-generated reasoning for assessment. The quality of responses is evaluated based on established criteria: Fluency and Relevance, rated on a five-point scale from unacceptable to excellent. Additionally, we assess the informativeness of the generated reasoning, considering whether it effectively includes key details, such as the speaker, the intended target, and the explanation for the sarcasm. The Informativeness metric spans from 0, indicating a lack of information, to 5, signifying highly informative reasoning.

**Results and Analysis**

In this section, we provide the results of all the three tasks followed by a comprehensive analysis of our approach and case studies.

**Main Results.** The most notable observation is the consistently substantial improvement demonstrated by SIRG across all metrics and tasks, encompassing sarcasm initiation (refer to Table 1 (TOP)), sarcasm detection (refer to Table 1 (TOP)), and sarcasm reasoning (refer to Table 1 (BOTTOM)). Examining the tables, specifically for the SIRD-English dataset and the sarcasm detection task, we
We provide a detailed analysis of Table 1: Experimental results for sarcasm initiation, detection, and reasoning on the SIRD dataset.

Table 2: Results of Human Evaluation on Sarcasm Reasoning.

Table 3: Task-wise analysis of English and Hinglish datasets.

Lastly, in terms of sarcasm reasoning generation, we observe a substantial improvement of 6.17% in terms of the F1 score compared to the baseline CMSEKI approach. In the context of sarcasm detection, it is evident that our SIRG methodology achieves a significant advancement over the CMSEKI baseline, achieving a difference of 0.04 in terms of the F1 score and 0.05 in terms of ROS score.

Table 4: Experimental results for sarcasm initiation, detection, and reasoning on the SIRD dataset.

In summary, the experimental results demonstrated that our SIRG approach consistently outperforms the existing baselines when it comes to generating appropriate responses across various manual evaluation metrics. The findings underscore the significance of a multifaceted approach in model training. It is evident that when distinct tasks are considered individually and in combination, the proposed model achieves a marked improvement in the model's efficiency. Additionally, the generated responses are not only fluent but also highly relevant to the given context, effectively encapsulating the essence of irony within the dialogue.

To evaluate the quality of the generated responses, we conducted human assessments based on standard evaluation metrics. The results indicate that our approach in model training provides comprehensive explanations for sarcasm.

Table 5: Model comparison for sarcasm reasoning.

Further analysis revealed that the SIRG model, a human-assessed approach using a randomly selected sample of 250 instances from the test set of a cross-validation fold, consistently outperforms the baselines across various manual evaluation metrics. The experimental results highlight the effectiveness of our approach in model training, thus providing comprehensive explanations for sarcasm.

Lastly, in terms of sarcasm reasoning generation, we observe a substantial improvement of 6.17% in terms of the F1 score compared to the baseline CMSEKI approach. In the context of sarcasm detection, it is evident that our SIRG methodology achieves a significant advancement over the CMSEKI baseline, achieving a difference of 0.04 in terms of the F1 score and 0.05 in terms of ROS score.

In summary, the experimental results demonstrated that our SIRG approach consistently outperforms the existing baselines when it comes to generating appropriate responses across various manual evaluation metrics. The findings underscore the significance of a multifaceted approach in model training. It is evident that when distinct tasks are considered individually and in combination, the proposed model achieves a marked improvement in the model's efficiency. Additionally, the generated responses are not only fluent but also highly relevant to the given context, effectively encapsulating the essence of irony within the dialogue.

To evaluate the quality of the generated responses, we conducted human assessments based on standard evaluation metrics. The results indicate that our approach in model training provides comprehensive explanations for sarcasm.
tasks are integrated, it provides a richer context, potentially aiding in better decision-making and inference by the model.

**Fusion at Different Layers.** In our approach, we combine both audio and video data using a fusion mechanism. This fusion takes place within the BART encoder, a component of our system. By experimenting with various layers of the encoder, we found that the most effective results are achieved when the fusion occurs before the final layer (layer 6). This choice yields the best outcomes, as demonstrated in the results presented in Table 5.

### Table 4: Ablation study for proposed SIRG where JF, ROS, BS, M, F1, and Acc denote the Jaccard F1, Recall-Oriented Score, BERTScore, METEOR, F1-score, and Accuracy, respectively.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Initiation</th>
<th>Reasoning</th>
<th>Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Hinglish</td>
<td>English</td>
</tr>
<tr>
<td></td>
<td>JF</td>
<td>ROS</td>
<td>JF</td>
</tr>
<tr>
<td>SIRG–clustering</td>
<td>0.778</td>
<td>0.789</td>
<td>0.731</td>
</tr>
<tr>
<td>SIRG–SF</td>
<td>0.767</td>
<td>0.0779</td>
<td>0.722</td>
</tr>
<tr>
<td>SIRG TV</td>
<td>0.721</td>
<td>0.786</td>
<td>0.726</td>
</tr>
<tr>
<td>SIRG TV</td>
<td>0.757</td>
<td>0.773</td>
<td>0.718</td>
</tr>
</tbody>
</table>

Table 5: Results of different fusion layers within BART

<table>
<thead>
<tr>
<th>Layer</th>
<th>English</th>
<th>Hinglish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ROS</td>
<td>ROUGE-L</td>
</tr>
<tr>
<td>1</td>
<td>0.803</td>
<td>0.774</td>
</tr>
<tr>
<td>2</td>
<td>0.804</td>
<td>0.775</td>
</tr>
<tr>
<td>3</td>
<td>0.803</td>
<td>0.771</td>
</tr>
<tr>
<td>4</td>
<td>0.799</td>
<td>0.773</td>
</tr>
<tr>
<td>5</td>
<td>0.799</td>
<td>0.771</td>
</tr>
<tr>
<td>6</td>
<td>0.81</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Figure 4: Case Study on SIRD dataset

**Ablation Study.** We further conducted an ablation study to demonstrate the effectiveness of our proposed model. In these experiments, we selectively omitted certain components, with the outcomes displayed in Table 4. Notably, we carried out two distinct tests: SIRG–clustering and SIRG–SF. When contrasted with the SIRG, both SIRG–clustering and SIRG–SF experienced a decrease across all metrics. It’s pertinent to point out that the values in parentheses indicate the change in scores compared to the main model. This analysis underscores the essential role of every component in determining the model’s comprehensive performance. The decline in metrics upon their removal underscores their contribution to the model’s robustness and accuracy.

**Case Study.** In Figure 4, we showcase case studies for both the English and Code-mixed (Hinglish) segments of the SIRD dataset in the context of the sarcasm reasoning task. The figure highlights that in the SIRD-English dataset, the reasoning generated by our proposed SIRG–TAVB framework exhibits higher accuracy, fluency, and information content when compared to the baseline MAF-TAVB approach, and it closely aligns with the actual ground-truth reasoning. The baseline approach tends to produce shorter reasoning, resulting in the omission of context and vital information. Likewise, in the case of the code-mixed SIRD dataset, it is evident that our proposed approach yields improved reasoning compared to the MAF-TAVB approach, and it is on par with the gold-standard reasoning provided for the given dialogue instance.

**Conclusion**

In this paper, we delved into the discourse structure of conversations infused with sarcasm and introduce a novel task - Sarcasm Initiation and Reasoning in Conversations (SIRC) - which is embedded in a multimodal environment and involves a combination of both English and code-mixed (Hinglish) interactions, the objective of the task is to discern the trigger or starting point of sarcasm. Additionally, the task involves producing a natural language explanation that rationalizes the satirical dialogues. To achieve this, we first introduced Sarcasm Initiation and Reasoning Dataset (SIRD) to facilitate our task and provide sarcasm initiation annotations and reasoning. We then developed a comprehensive model named Sarcasm Initiation and Reasoning Generation (SIRG), which is designed to encompass textual, audio, and visual representations. Our experimental outcomes, conducted on the SIRD dataset, demonstrate that our proposed framework established a new benchmark for both sarcasm initiation and its reasoning generation in the context of multimodal conversations.
References


Balduzzi, D.; Frean, M.; Leary, L.; Lewis, J.; Ma, K. W.-D.; and McWilliams, B. 2017. The shattered gradients problem: If resnets are the answer, then what is the question? In International Conference on Machine Learning, 342–350. PMLR.


