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

Gopendra Vikram Singh[1*], Mauajama Firdaus[2*], Dushyant Singh Chauhan[1*], Asif Ekbal[1], Pushpak Bhattacharyya[3]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]Department of Computing Science, University of Alberta, Canada
[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 https://www.iitp.ac.in/~ai-nlp-ml/resources.html#sarcasm-explain and 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-

*These authors contributed equally jointly serving as the primary authors.

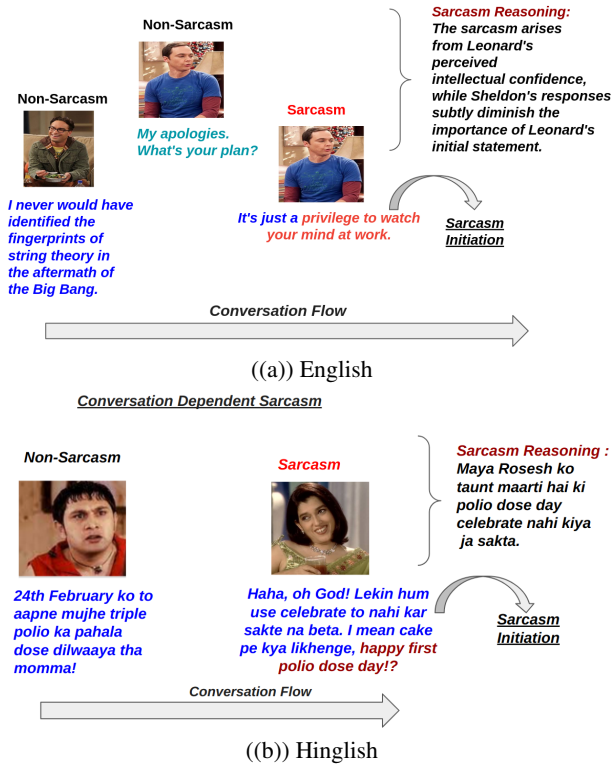

Figure 1: Example of SIRD-English and SIRD-Hinglish datasets showing sarcasm Initiation and Reasoning

ple sentences or the entire conversation context. The cause or span of sarcasm depends on linguistic and contextual clues, including tone of voice, exaggerated expressions, irony, or contrasting meanings. As depicted in Figure 1, the portion denoted in red indicates the sarcasm initiation span, serving as the discernible trigger for the sarcasm. This marked segment encapsulates the underlying cause that sparks the sarcastic tone. Moreover, a well-constructed reasoning, providing contextual details, insight into the speaker's perspective, and the intended target of the sarcasm, accompanies this initiation span.

Motivated by the intricate nature of the challenge at hand, we embark on a pioneering endeavor within this paper. We introduce an innovative and multifaceted task that transcends conventional sarcasm detection. Alongside identifying instances of sarcasm within a dialogue, we delve deeper to unearth the instigating factors behind the initiation of sarcasm, all while accommodating both the English and code-mixed linguistic contexts. This novel task is aptly labeled as Sarcasm Initiation and Reasoning in Conversations (SIRC), underscoring our pursuit of comprehensively understanding the dynamics of sarcasm within conversational exchanges.

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, Bhattacharyya, 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.

After an extensive literature review, it becomes evident that the adoption of a multimodal approach in sarcasm detection (Bedi et al. 2021; Li et al. 2021; Wang et al. 2022; Chauhan et al. 2020b; Liang et al. 2022; Chauhan et al. 2022) is a relatively recent strategy, diverging from conventional text-based classification methods. 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.

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, Desai, Chakraborty, and Akhtar (2022) bring to the fore an innovative challenge named Multimodal Sarcasm Explanation (MuSE), revolving around generating a natural language explanation for a multimodal sarcastic expression encompassing both an image and a caption.

In (Desai, Chakraborty, and Akhtar 2022), the authors introduce a new task known as Multimodal Sarcasm Explanation (MuSE). In this problem, the objective is to provide a natural language explanation for a multimodal sarcastic post, which comprises both an image and a caption. Similarly, in (Kumar et al. 2022), the authors study the discourse structure of conversations characterized by sarcasm. They introduce a new task titled Sarcasm Explanation in Dialogue (SED), which operates within a context that encompasses both multimodal elements and a mixture of languages. This task is geared towards generating natural language explanations for satirical conversations, elucidating the intricacies of the exchanged content. Our present research stands apart

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.

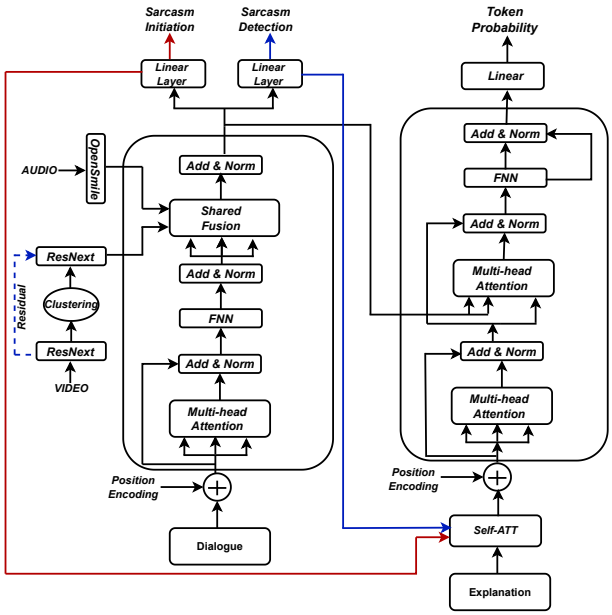

Figure 2: Overall architecture

Video Encoder. The visual world and facial expressions offer rich emotional cues. The facial emotions and visual environment from the utterance video are captured using 3D-ResNeXt[1] (Hara, Kataoka, and Satoh 2018), which delivers 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.

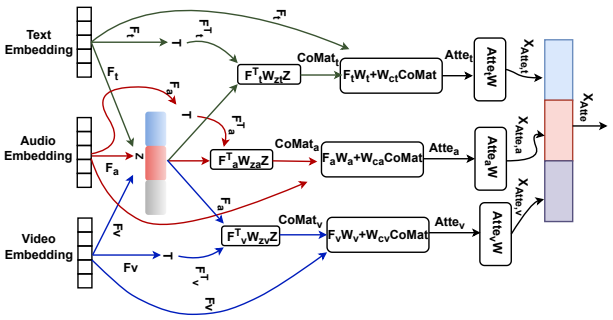

Figure 3: Shared-Fusion between the textual, audio and visual modalities

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 mecha-

[1]https://github.com/kaiqiangh/extracting-video-features-ResNeXt

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: $[F_t; F_t; F_t] \in \mathcal{R}^{D*L}$, where D denoted by $D = x_t + x_a + x_v$ shows the concatenated features'(text, audio, video) dimension. For the given multimodal utterance (M_u), 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:

$$CoMat_t = tanh\frac{F_t^T W_{ztZ}}{\sqrt{x}}$$

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

$$CoMat_a = tanh\frac{F_a^T W_{zaZ}}{\sqrt{x}}$$

$$CoMat_v = tanh\frac{F_v^T W_{zvZ}}{\sqrt{x}}$$

The shared correlation matrices $CoMat_t$, $CoMat_v$, and $CoMat_a$ 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 $CoMat_t$, $CoMat_v$, and $CoMat_a$. In order to improve the performance of the system, the suggested strategy effectively takes advantage of the complimentary nature of 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_{ct} 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:

$$Atte_t = ReLu(W_t F_t + W_{ct} CoMat_t^T)$$

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

$$Atte_v = ReLu(W_v F_v + W_{cv} CoMat_v^T)$$

$$Atte_a = ReLu(W_a F_t + W_{ca} CoMat_a^T)$$

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

$$X_{Atte_t,t} = WAtte_t + F_t$$
$$X_{Atte_a,a} = WAtte_a + F_a$$
$$X_{Atte_v,v} = WAtte_v + F_v$$

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

$$X_{Atte} = [X_{Atte_t,t}; X_{Atte_a,a}; X_{Atte_v,v}]$$

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

$$\mathcal{L} = \mathcal{L}_{SD} + \mathcal{L}_{SI} + \mathcal{L}_{SR} \quad (1)$$

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 v/s 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 $\mathcal{D}$ 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$_M$ and 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

MODEL	Initiation										Detection			
	English					Hinglish					English		Hinglish	
	FM	PM	HD	JF	ROS	FM	PM	HD	JF	ROS	F1	Acc	F1	Acc
BiRNN-Attn	24.65	18.19	0.47	0.64	0.69	18.86	13.54	0.38	0.57	0.60	75.76	77.75	57.13	59.22
CNN-GRU	23.74	17.32	0.45	0.65	0.70	17.54	14.32	0.37	0.60	0.61	74.79	76.24	57.97	60.33
BiRNN-HateXplain	29.31	22.77	0.53	0.70	0.72	19.21	16.34	0.41	0.61	0.62	76.23	76.88	59.47	62.62
SpanBERT	31.24	24.23	0.55	0.72	0.73	21.11	20.16	0.43	0.64	0.63	75.17	77.62	61.22	63.66
BERT-HateXplain	30.77	23.43	0.57	0.72	0.75	22.14	19.54	0.45	0.65	0.63	76.75	78.49	62.48	65.68
CMSEKI	33.65	24.32	0.56	0.75	0.76	24.73	23.87	0.47	0.70	0.73	77.21	79.74	64.52	67.67
SIRG$_M$	**35.62**	**27.24**	**0.62**	**0.78**	**0.79**	**29.19**	**25.77**	**0.50**	**0.74**	**0.76**	**82.24**	**84.71**	**68.57**	**71.74**
SIRG$_B$	**37.12**	**28.33**	**0.64**	**0.79**	**0.81**	**31.27**	**27.22**	**0.52**	**0.75**	**0.78**	**84.11**	**85.91**	**69.99**	**72.13**

MODEL	Experimental results: Reasoning									
	English					Hinglish				
	Rouge-L	BS	M	BLEU-3	BLEU-4	Rouge-L	BS	M	Bleu3	Bleu4
PGN	20.44	71.69	25.72	3.98	1.68	17.46	71.90	23.54	1.58	0.52
mBART	38.17	75.99	31.78	12.76	9.34	31.50	73.83	21.03	6.07	3.39
BART	40.33	76.43	33.24	14.54	11.11	33.49	76.03	26.65	5.96	2.89
MAF-TAV$_M$	43.19	77.44	36.19	17.44	16.79	36.60	76.70	27.42	9.78	5.74
MAF-TAV$_B$	45.26	79.52	37.44	19.61	17.23	37.37	77.67	30.40	12.37	8.57
SIRG-TAV$_M$	49.61	84.77	40.31	24.70	23.72	41.21	82.34	35.33	15.99	12.68
SIRG-TAV$_B$	52.22	86.11	42.26	27.21	25.13	42.81	83.23	36.48	17.66	16.32

Table 1: Experimental results for sarcasm initiation, detection, and reasoning on the SIRD dataset; **TOP**:: Experimental results for sarcasm initiation and detection where FM, PM, HD, JF, ROS, F1, and Acc denote the F-Measure-Modified, Precision-Modified, Hamming Distance, Jaccard F1, Recall-Oriented Score, F1-score, and Accuracy, respectively; **BOTTOM**:: Experimental results for sarcasm reasoning where BS and M denote the BERTScore and METEOR

achieve a substantial improvement of 6.17% in terms of the *F1 score* compared to the baseline *CMSEKI* approach. In the context of sarcasm initiation, it is evident that our *SIRG* methodology achieves a significant advancement over the *CMSEKI* baseline, achieving a difference of 0.04 in terms of *JF score* and 0.05 in terms of *ROS score*.

Lastly, in terms of sarcasm reasoning generation, we observe significant enhancements of 7.9 and 6.96 in comparison to the baseline *MAF*$_B$ approach, as indicated by the improvements in BLEU-4 and Rouge-L scores, respectively. Likewise, we also observe a substantial increase of 6.59 in the BERTScore. In addition, when considering the SIRD-Hinglish dataset, we consistently witness improvements in performance across all three tasks. Consequently, we can confidently assert that our proposed approach, *SIRG*$_B$, stands out as the most effective solution for all three tasks based on standard evaluation metrics.

Human Evaluation. To evaluate the quality of the generated utterances by the *SIRG* model, a human assessment was conducted using a randomly selected sample of 250 instances from the test set of a cross-validation fold. In alignment with the experimental findings (refer to Table 1), the outcomes of the human evaluation (see Table 2) corroborate the superior performance of *SIRG* compared to the existing baselines when it comes to generating appropriate reasoning. It is evident that *SIRG* consistently outperforms the baselines across various manual evaluation metrics. The generated responses are not only fluent but also highly relevant to the given context, effectively encapsulating crucial information including the speaker's perspective, the intended target, and the essence of irony within the dialogue, thus providing comprehensive explanations for sarcasm.

Task-wise Analysis. We provide a detailed analysis of each task for both datasets in Table 3. We'd like to highlight that I+R+D signifies the most effective model, namely, $SIRG\text{-}TAV_B$. By incorporating these specific tasks — initiation (I), reasoning (R), and detection (D) — there's a marked improvement in the model's efficiency. Additionally, these findings underscore the significance of a multifaceted approach in model training. It's evident that when distinct

Models	Fluency	Adequacy	Informativeness
MAF-TAV$_B$	3.08	3.11	2.97
SIRG$_B$	**3.21**	**3.29**	**3.31**
MAF-TAV$_M$	3.16	3.05	3.11
SIRG$_M$	**3.36**	**3.41**	**3.83**

Table 2: Results of Human Evaluation on Sarcasm Reasoning task

Setup	English			Hinglish		
	BS	ROS	F1	BS	ROS	F1
I	-	77	-	-	75	-
R	81.13	-	-	79.21	-	-
D	-	-	80.97	-	-	65.94
I+R	84.17	79	-	82.11	77	-
I+D	-	78	82.45	-	77	67.37
R+D	83.79	-	82.13	81.41	-	67.66
I+R+D	**86.11**	**81**	**84.11**	**83.23**	**78**	**69.99**

Table 3: Task-wise analysis of English and Hinglish datasets where I: Sarcasm Initiation, R: Sarcasm Reasoning, D: Sarcasm Detection and I+R+D denotes the best model, i.e., SIRG-TAV$_B$ (c.f. Table 1)

Setup	Initiation				Reasoning				Detection			
	English		*Hinglish*		*English*		*Hinglish*		*English*		*Hinglish*	
	JF	*ROS*	*JF*	*ROS*	*BS*	*M*	*BS*	*M*	*F1*	*Acc*	*F1*	*Acc*
SIRG–clustering	0.778	0.789	0.731	0.75	83.98	40.49	80.79	34.37	81.12	82.23	67.86	69.69
SIRG–SF	0.767	0.0779	0.722	0.748	83.24	40.24	80.5	34.04	81.10	83.10	67.1	69.34
SIRG$_{TA}$	0.721	0.786	0.726	0.758	83.13	39.20	79.24	34.20	81.23	82.89	67.43	69.36
SIRG$_{TV}$	0.757	0.773	0.718	0.749	82.45	38.51	79.35	33.29	80.61	82.58	66.70	68.94

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

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.

Layer	*English*		*Hinglish*	
	ROS	*ROUGE-L*	*ROS*	*ROUGE-L*
1	0.803	51.77	0.774	41.44
2	0.804	51.29	0.775	41.24
3	0.803	51.46	0.771	41.59
4	0.799	51.69	0.773	41.31
5	0.799	51.07	0.771	41.57
6	***0.81***	***52.22***	***0.78***	***42.81***

Table 5: Results of different fusion layers within BART

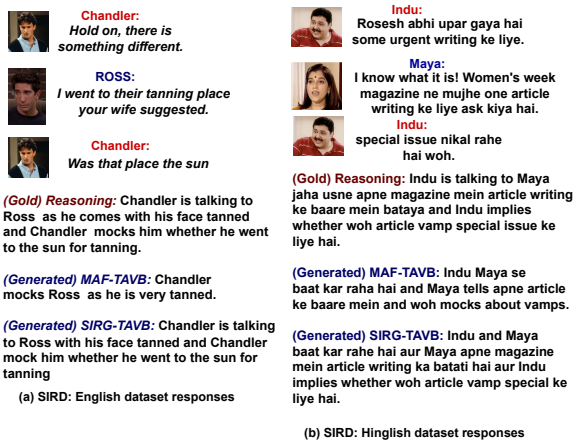

Figure 4: Case Study on SIRD dataset

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-TAV*$_B$ framework exhibits higher accuracy, fluency, and information content when compared to the baseline *MAF-TAV*$_B$ 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-TAV*$_B$ 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

Babanejad, N.; Davoudi, H.; An, A.; and Papagelis, M. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th international conference on computational linguistics*, 225–243.

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.

Bedi, M.; Kumar, S.; Akhtar, M. S.; and Chakraborty, T. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*.

Campbell, J. D.; and Katz, A. N. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6): 459–480.

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Chauhan, D. S.; Dhanush, S.; Ekbal, A.; and Bhattacharyya, P. 2020a. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, 281–290.

Chauhan, D. S.; Dhanush, S.; Ekbal, A.; and Bhattacharyya, P. 2020b. 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*, 4351–4360.

Chauhan, D. S.; Singh, G. V.; Arora, A.; Ekbal, A.; and Bhattacharyya, P. 2022. An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems*, 257: 109924.

Desai, P.; Chakraborty, T.; and Akhtar, M. S. 2022. Nice perfume. How long did you marinate in it? Multimodal Sarcasm Explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10563–10571.

Eyben, F.; Wöllmer, M.; and Schuller, B. W. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, 1459–1462. ACM.

Ghosh, S.; Ekbal, A.; and Bhattacharyya, P. 2022. A Multitask Framework to Detect Depression, Sentiment and Multilabel Emotion from Suicide Notes. *Cogn. Comput.*, 14(1): 110–129.

Ghosh, S.; Roy, S.; Ekbal, A.; and Bhattacharyya, P. 2022. CARES: CAuse Recognition for Emotion in Suicide Notes. In *European Conference on Information Retrieval*, 128–136. Springer.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 6546–6555. Computer Vision Foundation / IEEE Computer Society.

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

Joshi, A.; Sharma, V.; and Bhattacharyya, P. 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)*, 757–762.

Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kumar, S.; Kulkarni, A.; Akhtar, M. S.; and Chakraborty, T. 2022. When did you become so smart, oh wise one?! Sarcasm Explanation in Multi-modal Multi-party Dialogues. *arXiv preprint arXiv:2203.06419*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, J.; Pan, H.; Lin, Z.; Fu, P.; and Wang, W. 2021. Sarcasm detection with commonsense knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3192–3201.

Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; and Xu, R. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1767–1777.

Liu, B.; and Lane, I. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14867–14875.

Olkoniemi, H.; Ranta, H.; and Kaakinen, J. K. 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.

Poria, S.; Cambria, E.; Hazarika, D.; and Vij, P. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Poria, S.; Majumder, N.; Hazarika, D.; Ghosal, D.; Bhardwaj, R.; Jian, S. Y. B.; Hong, P.; Ghosh, R.; Roy, A.; Chhaya, N.; et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5): 1317–1332.

Schifanella, R.; de Juan, P.; Tetreault, J.; and Cao, L. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, 1136–1145. ACM.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Veale, T.; and Hao, Y. 2010. Detecting ironic intent in creative comparisons. In *ECAI 2010*, 765–770. IOS Press.

Wang, J.; Sun, L.; Liu, Y.; Shao, M.; and Zheng, Z. 2022. Multimodal Sarcasm Target Identification in Tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8164–8175.

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

Zhang, Z.; Robinson, D.; and Tepper, J. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, 745–760. Springer.