Explaining (Sarcastic) Utterances to Enhance Affect Understanding in Multimodal Dialogues

Shivani Kumar¹, Ishani Mondal², Md Shad Akhtar¹, Tanmoy Chakraborty³

¹Indraprastha Institute of Information Technology Delhi, India
²University of Maryland, College Park
³Indian Institute of Technology Delhi, India

shivaniku@iiitd.ac.in, ishani340@gmail.com, shad.akhtar@iiitd.ac.in, tanchak@iitd.ac.in

Abstract

Conversations emerge as the primary media for exchanging ideas. Identifying various affective qualities, such as sarcasm, humour, and emotions, is paramount for comprehending the true connotation of the emitted utterance. However, one of the major hurdles faced in learning these affect dimensions is the presence of figurative language viz. irony, metaphor, or sarcasm. We hypothesize that any detection system constituting the exhaustive and explicit presentation of the emitted utterance would improve the overall comprehension of the dialogue. To this end, we explore the task of Sarcasm Explanation in Dialogues that aims to unfold the hidden irony behind sarcastic utterances. We propose MOSES, a deep neural network, which takes a multimodal (sarcastic) dialogue instance as an input and generates a natural language sentence as its explanation. Subsequently, we leverage the generated explanation for various natural language understanding tasks in a conversational dialogue setup, such as sarcasm detection, humour identification, and emotion recognition. Our evaluation shows that MOSES outperforms the state-of-the-art system for SED by an average of ~ 2% on different evaluation metrics, such as ROUGE, BLEU, and METEOR. Further, we observe that leveraging the generated explanation advances three downstream tasks for affect classification - an average improvement of ~ 14% F1-score in the sarcasm detection task and ~ 2% in the humour identification and emotion recognition task. We also perform extensive analyses to assess the quality of the results.

Introduction

Expressing oneself eloquently to our conversation partner requires employing multiple affective components such as emotion, humour, and sarcasm. All such attributes interact with each another to present a concrete definition of an uttered statement (Roberts and Kreuz 1994). While affects such as emotion and humour deem easier to comprehend, sarcasm, on the other hand, is a challenging aspect to comprehend (Olkoniemi, Ranta, and Kaakinen 2016). Consequently, it becomes imperative for NLP systems to capture and understand sarcasm in its entirety. Sarcasm Explanation in Dialogues (SED) is a new task proposed recently in this direction (Kumar et al. 2022). In this work, we scour the task of SED and considerably improve the performance by proposing MOSES, a deep neural network which leverages multimodal knowledge such as audio and video can assist in comprehending sarcasm. The identical text, “This party is so fun”, can result in diverse classes when integrated with different multimodal signals. For instance, when said with a neutral tone and a disgusted face (Case II), it results in a ‘sarcastic’ verdict. On the other hand, when the same utterance is expressed with a low tone and while the speaker is dancing (Case III), we can assume the speaker means what they say without any sarcasm. To capture multimodality efficaciously, it is vital to grant a prime prerogative to each input modality in order to capture its peculiarities. To this end, we propose a spotlight-aware fusion mechanism, where the final multimodal amalgamated vector is tailored by paying special attention to individual modalities.

All affective components, such as sarcasm, humour, and emotion, work in tandem to convey a statement’s intended meaning (Hasan et al. 2021; Chauhan et al. 2020). Accordingly, we hypothesize that understanding one of the affective markers, like sarcasm, in its entirety will influence comprehending others. Consequently, in this work, we deal with leveraging sarcasm explanations for three affect understanding tasks in dialogues, namely sarcasm detection, humour identification, and emotion recognition. The performance obtained from these tasks can be employed as a method to estimate the relevance of the SED task extrinsically.

We summarize our contributions below:
1. We explore the new task of SED and propose a novel model, MOSES, for it.
2. We compare MOSES with existing baselines and obtain state-of-the-art results for the SED task.
3. We show the application of the generated explanations in understanding different affective components – sarcasm, emotions and humour.
4. We show extensive quantitative and qualitative studies for all our experiments.

Reproducibility: The source code for MOSES and the execution instructions are present here: https://github.com/LCS2-IIITD/MOSES.git.

Related Work

Sarcasm. Figurative language such as sarcasm plays an integral role in resolving the veiled meaning of an uttered statement. Earlier studies dealt with sarcasm identification in standalone texts like tweets and reviews. (Kreuz and Caucci 2007; Tsur, Davidov, and Rappoport 2010; Joshi, Sharma, and Bhattacharyya 2015; Peled and Reichart 2017). A detailed summary of these studies can be found in the survey compiled by Joshi, Bhattacharyya, and Carman (2017). Several work explored sarcasm in other languages such as Hindi (Bharti, Sathy Babu, and Jena 2017), Arabic (Abu Farha and Magdy 2020), Spanish (Ortega-Bueno et al. 2019), Italian (Cignarella et al. 2018), or even code-mixed (Swami et al. 2018) languages.

Sarcasm and Dialogues. Linguistic and lexical traits were the primary sources of sarcasm markers in previous investigations (Kreuz and Caucci 2007; Tsur, Davidov, and Rappoport 2010). However, in more contemporary studies, attention-based approaches are used to capture the inter- and intra-sentence interactions in the text (Tay et al. 2018; Xiong et al. 2019; Srivastava et al. 2020). In terms of conversations, Ghosh, Richard Fabbri, and Muresan (2017) harnessed attention-based RNNs to capture context and determine sarcasm.

Sarcasm and Multimodality. Castro et al. (2019) proposed a multimodal, multiparty, English dataset called MUSTARD to benchmark the task of multimodal sarcasm identification in conversation. Subsequently, Chauhan et al. (2020) devised a multi-task framework by leveraging interdependency between emotions and sarcasm to solve the task of multimodal sarcasm detection. Another work (Hasan et al. 2021) established the interdependency of humour with sarcasm by suggesting a humour knowledge enriched Transformer model for sarcasm detection. In the code-mixed scenario, Bedi et al. (2021) proposed MASAC, a multimodal, multiparty, code-mixed dialogue dataset for humour and sarcasm detection. In the bimodal setting, sarcasm identification with tweets containing images has also been well explored (Cai, Cai, and Wan 2019; Xu, Zeng, and Mao 2020; Pan et al. 2020).

Beyond Sarcasm Detection. Sarcasm generation is another direction that practitioners are inquisitive about due to its forthright benefit in enhancing chatbot engagement. Thereby, Mishra, Tater, and Sankaranarayanan (2019) induced sarcastic utterances by presenting context incongruity through fact removal and incongruous phrase insertion. A retrieve-and-edit-based unsupervised approach for generating sarcasm was proposed by Chakrabarty et al. (2020) that exploits semantic incongruity and valence reversal to convert non-sarcastic instances to sarcastic ones. On the other hand, while detecting irony is crucial, it is insufficient to capture the cardinal connotation of the statement. Consequently, Dubey, Joshi, and Bhattacharyya (2019) examined the task of converting sarcastic utterances into their non-sarcastic counterparts using deep neural networks.

In this work, we explore the task of Sarcasm Explanation in Dialogues, the second attempt after Kumar et al. (2022). SED aims to generate natural language explanations for a disseminated multimodal sarcastic conversation. We present a new model, MOSES, which enhances the current state-of-the-art for the SED task. However, unlike Kumar et al. (2022), we perform both intrinsic and extrinsic evaluations to show the efficacy and usefulness of our model. We leverage the generated explanations to improve three affect understanding tasks – sarcasm detection, humour identification, and emotion recognition in dialogues.

Dataset

Human conversations often take place employing a variety of languages. The phenomenon of using a blend of more than one language to communicate is dubbed code-mixing. Due to the prevalence of code-mixing in today’s world, we consider the WITS dataset (Kumar et al. 2022), which contains code-mixed dialogues (English-Hindi) from an Indian TV series. The dataset comprises multimodal, multiparty, code-mixed, sarcastic conversations where each sarcastic instance is annotated with a corresponding natural language code-mixed explanation.

In order to gauge the effect of sarcasm explanation on affective attributes, we augment the WITS dataset to perform sarcasm detection, humour identification, and emotion recognition on it. We create instances for sarcastic and non-sarcastic utterances with their context to perform sarcasm detection. We call this variation of the dataset sWITS. Adapted from MASAC (Bedi et al. 2021), WITS can also be mapped to annotations for humour identification, where each utterance contains a binary marker showcasing whether the utterance is amusing or not. Consequently, we map each instance in sWITS to its corresponding humour annotation.

Additionally, we determine emotion labels for the instances at hand and identify the following emotions – sadness, joy, anger, and neutral. Three annotators were involved in this phase and achieved an inter-annotator agreement of 0.86. More information can be found in the supplementary. Accordingly, we obtain four variations of the dataset:

1. WITS: It contains multimodal, multiparty, code-mixed, sarcastic instances with associated explanations.
2. sWITS: It contains sarcastic and non-sarcastic instances constructed from WITS. The last utterance of each instance is marked by a binary tag indicating whether the statement contains sarcasm or not.
3. hWITS: For each instance created in sWITS, each target utterance is marked with another binary label revealing the existence of humour in it.
Table 1: Statistics of the sarcasm, humour, and emotion (N: Ntrl: Neutral, Ang: Anger) datasets in consideration (number of dialogue instances marked as sarcastic (#S), non-sarcastic (#NS), non-humorous (#NH), and humorous (#H)).

<table>
<thead>
<tr>
<th></th>
<th>WITS</th>
<th>eWITS</th>
<th>hWITS</th>
<th>eWITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>#S</td>
<td>1792</td>
<td>1669</td>
<td>1792</td>
<td>1590</td>
</tr>
<tr>
<td>#NS</td>
<td>1792</td>
<td>2795</td>
<td>995</td>
<td>1147</td>
</tr>
<tr>
<td>#S</td>
<td>218</td>
<td>362</td>
<td>112</td>
<td>196</td>
</tr>
<tr>
<td>#S</td>
<td>224</td>
<td>367</td>
<td>106</td>
<td>195</td>
</tr>
<tr>
<td>#S</td>
<td>224</td>
<td>196</td>
<td>87</td>
<td>141</td>
</tr>
<tr>
<td>#S</td>
<td>224</td>
<td>131</td>
<td>57</td>
<td>70</td>
</tr>
<tr>
<td>#S</td>
<td>224</td>
<td>141</td>
<td>70</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 1 illustrates the elementary statistics for the explained dataset variations. More details about the dataset are present in the supplementary.

Proposed Method

This section illustrates the working of our proposed model, MOSES as presented in Figure 2. The existing SED model, MAF (Kumar et al. 2022), which uses a modified version of context-aware attention (Yang et al. 2019), takes the multimodal (audio/video) vectors as context and fuses them with the text modality to generate multimodal fused text vectors. This way of multimodal fusion makes text the primary modality and treats the other signals (acoustic and visual) as secondary. Such a fusion technique might result in the downplay of the audio and video modalities. However, in the complete duration of the discourse, modalities other than text could play the deciding role in resolving the affects in consideration. Consequently, we propose using context-aware fusion in such a way that each modality gets a chance to play a pivotal role in the fusion computation.

The existing MAF module consists of an adapter-based module comprising two modules. The two modules – Multimodal Context Aware Attention (MCA2) and Global Information Fusion (GIF) together make up the Multimodal Aware Fusion (MAF) module. Given the three input signals, namely text, audio, and video, the MCA2 module effectively introduces multimodal information in the textual representations. Further, the GIF module combines the multimodal infused textual representations. We insert another module in the pipeline, Modality Spotlight (MS), which is responsible for attending to each modality by treating it as the primary modality and the rest as the context. We explain each of these modules below.

Multimodal Context-Aware Attention (MCA2). The textual modality directly interacts with the other modalities in the standard fusion scheme, which uses dot-product based cross-modal attention. The multimodal representation acts as the key and value vectors while the text serves as the query. However, such a direct fusion of multimodal information from different embedding subspaces can lead to an inefficient representation that cannot capture contextual information. Consequently, inspired by Yang et al. (2019), a context-aware attention block is used instead of dot product attention in the MCA2 module.

The multimodal context-aware attention first generates multimodal information conditioned key and value vectors using a primary representation, $H$ which depends on the choice of dominant modality in consideration. For example, $H$ can be obtained from any language model, such as BART (Lewis et al. 2020), if text is to be considered as the primary modality. To generate information fused vectors, the MCA2 module first needs to convert $H$ into key, query, and value vectors, $q, k, v \in \mathbb{R}^{n \times d}$, respectively, as illustrated in Equation 1. Here, $W_q$, $W_k$, and $W_v \in \mathbb{R}^{d \times d}$ are learnable parameters where the input, of maximum length $n$, is represented by a vector of dimension $d$.

$$[qkv] = H[W_qW_kW_v]$$

If we consider $M \in \mathbb{R}^{n \times d}$ a multimodal vector, the multimodal information fused key and value vectors $k_m, v_m \in \mathbb{R}^{n \times d}$, are generated following Equation 2. Here, $\lambda \in \mathbb{R}^{n \times 1}$ decides the amount of multimodal information to capture into the primary modality.

$$[k_m \ v_m] = (1 - [\lambda_k \ \lambda_v]) [k \ v] + [\lambda_k \ \lambda_v] (M [U_k \ U_v])$$

The parameter $\lambda$ is learned using a gating mechanism, as shown in Equation 3. Note that the matrices, $U_k, U_v \in \mathbb{R}^{d \times d}$, and $W_{k1}, W_{v1}, W_{k2}, W_{v2} \in \mathbb{R}^{d \times 1}$, in Equations 2 and 3 are trained with the model.

$$[\lambda_k \ \lambda_v] = \sigma([k \ v] W_{k1} + M [U_k \ U_v] W_{v1} + W_{k2} + W_{v2})$$

Once the modified key and value pair is obtained, the traditional dot-product attention is used to obtain the final multimodal fused vectors.

Modality Spotlight (MS). We discussed how we can generate multimodal infused vector representation considering one modality as primary and the rest as context. Our work deals with three modalities – text, acoustic, and visual. The spotlight module is responsible for treating each of these modalities as the primary modality at a time and generating the corresponding fused vectors. For instance, if text is considered the primary modality, then we need to calculate two multimodal fused vectors, $H_{Ta}$ and $H_{Tv}$, such as audio and video, play the role of context in the representations, respectively. Similarly, when audio and video are considered the primary source of information, $H_{TA}$ and $H_{TV}$ are calculated. Note that we do not calculate $H_{AV}$ because we are dealing with a textual generation task where the textual information plays the preliminary role.

Apart from bi-modal interactions, we also deal with tri-modal interactions in our work, where all three modalities are infused using the GIF module. Unlike bi-modal fusion, it is unfair to let text be the only primary modality in the tri-modal fusion. Consequently, we compute three tri-modal vectors, $H_{TaV}, H_{tAV}$, and $H_{tAV}$, such that text, audio, and video individually play the primary role, respectively.
Global Information Fusion (GIF). The GIF module is responsible for combining the information from multiple modalities together in an efficient manner. $G$ gates are used to control the amount of information disseminated by each modality, where $2 \leq G \leq 3$ is the number of modalities to fuse. For instance, if we calculate the interaction between the text and audio modalities with text being the primary source of information, we will first need to calculate the gated information from the audio representation using Equation 4.

$$g_a = [H \oplus H_a]W_a + b_a$$  \hspace{1cm} (4)

where $W_a$ and $b_a$ are learnable matrices, and $\oplus$ denotes vector concatenation. The final representation to be passed on to the next encoder layer will be obtained using Equation 5.

$$H_{Ta} = H + g_a \odot H_a$$  \hspace{1cm} (5)

On similar lines, if we are to calculate the tri-modal representation keeping the text as the primary modality, we first compute the gated vector for audio and video and then compute a weighted combination of the three modalities. The following sequence of equations illustrates this process.

$$g_a = [H \oplus H_a]W_a + b_a$$

$$g_v = [H \oplus H_v]W_v + b_v$$

$$H_{Tav} = H + g_a \odot H_a + g_v \odot H_v$$

Likewise, we calculate the following set of vectors: $H_{Ta}$, $H_{Tv}$, $H_{TvAv}$, $H_{Tav}$, and $H_{tav}$. Further, another GIF module is used to conglomerate these seven vectors, as shown in Equation 6.

$$H_{all} = g_t \odot H + g_{Ta} \odot H_{Ta} + g_{Tv} \odot H_{Tv} + g_{TvAv} \odot H_{TvAv} + g_{tav} \odot H_{tav}$$

$$+ g_{1Av} \odot H_{1Av} + g_{tav} \odot H_{tav}$$  \hspace{1cm} (6)

Experiment and Results

This section illustrates the feature extraction strategy we use and the baseline systems to which we compare our model, followed by the results we obtain for the SED task. We use the standard generative metrics – ROUGE-1/2/L (Lin 2004), BLEU-1/2/3/4 (Papineni et al. 2002), and METEOR (Denkowski and Lavie 2014) to capture the syntactic and semantic performance of our system. Details about the execution process and the hyperparameters used are mentioned in the supplementary.

Feature Extraction

The primary challenges for generating vector representations for the instances in WITS come from the code-mixed and multimodal aspects of the dataset. We alleviate these by proposing intelligent feature extraction methods.

Text: The textual input in WITS is present in romanised code-mixed format. Thereby, it may contain terms with the same meaning but varying spellings that are phonetically identical. For instance, the word “main” in Hindi (translating to “I” in English) can be written as “main” or “mein”. To capture the similarity between all these spelling variations, we propose using Pronunciation Embeddings (PE) that capture the phonetic equivalence between the words of the input text. We convert the text into a standard speech format using python’s gTTS library\(^1\). This converted audio does not contain any tone or pitch variation for any term and thus, sounds the same for phonetically similar terms. We then extract the audio features from this converted speech using the method described below. This pronunciation vector is fused with the text representation, obtained from any encoder model like BART, using the GIF module to obtain the final text representation.

Audio: We use 154 dimension features capturing the loudness, and Mel Frequency Cepstral Coefficient (MFCCs) for each instance using the eGeMAPS model (Eyben et al. 2016). A Transformer encoder (Vaswani et al. 2017) is used for further processing of these features.

Video: These features are obtained using ResNext-101 (Hara, Kataoka, and Satoh 2018). A resolution of 720 pixels, a window length of 16, and a frame rate of 1.5 are used to obtain 2048 dimensional visual features. Analogous to the acoustic representation, a Transformer encoder captures the sequential conversation context in the resultant vectors.

Comparative Systems

We use various established sequence-to-sequence (seq2seq) models to obtain the most promising textual representations for the discourse. RNN: The openNMT\(^2\) implementation of

\(^{1}\)https://pypi.org/project/gTTS/
\(^{2}\)https://github.com/OpenNMT/OpenNMT-py
Table 2: Experimental results (Abbreviation: R1/2/L: ROUGE1/2/L; B1/2/3/4: BLEU1/2/3/4; M: METEOR; PGN: Pointer Generator Network). Final row denotes MOSES including the pronunciation and spotlight modules.

Table 3: Ablation results on MOSES (Abbreviation: DPA: Dot Product Attention).

visual cues boosts the performance by around 1% across all metrics. This phenomenon can be attributed to the fact that audio alone may cause confusion while understanding sarcasm, and visual hints may help in such times (Case III in Figure 1). Thereby, improving the visual feature representations can be one of the future directions. Finally, when we add all multimodal signals together, we observe the best performance yet with an average increase of further around 1% across majority metrics.

Modality Spotlight: As hypothesised, we obtain the best performance for sarcasm understanding when all the three modalities are used in tandem. However, earlier methods for SED provided limelight to only textual representations (Kumar et al. 2022). We argue that especially in the case of sarcasm, multimodal signals such as audio and video might play the principal role in many instances. To comprehend the rotating importance of modalities, we use the spotlight module that aims to treat each modality as the primary modality while calculating the final representation. We observe an increase of around 2% across all evaluation metrics as shown in Table 2. These results directly support our hypothesis of the effect of multimodality in sarcasm analysis.

Ablation Study

To highlight the importance of all modules in consideration, we perform extensive ablation studies on the WITS dataset. Table 3 shows the results when we add the different proposed modules to our system sequentially. The first row highlights the BART model’s results for the text modality which results in a ROUGE-2 of 11.91%. As illustrated, the use of naive trimodal concatenation \((T \oplus A \oplus V)\) of text, audio, and video representations produces a noisy fusion resulting in decreased performance \((-10.2\% \text{ ROUGE-2})\). Next, we try with the standard dot-product attention, which, being a comparatively smarter way of multimodal fusion, results in a slightly improved performance over the text-only modality \((+2\% \text{ ROUGE-2})\). Further, adding the multimodal context-aware attention module (MCA2) and replacing standard dot-product attention, produces a further performance boost by around 1% across all metrics, signifying the importance of the intelligent fusion that the MCA2 module provides us. The performance is increased even more when the GIF module is introduced to compute the final multimodal vector representation \((+4\% \text{ ROUGE-2})\), signifying the positive effect gated fusion has on efficient multimodal representations. Next, we incorporate pronunciation embeddings (PE) into the model and observe another performance boost across majority metrics \((\sim 1\%)\), suggesting that we can obtain better code-mixed representations by reducing the spelling am-

---

1https://huggingface.co/facebook/mmbart-large-50-many-to-

many-mmt
Table 4: Generated samples from test set. The last utterance is the sarcastic utterance for each dialogue.

<table>
<thead>
<tr>
<th>Dialogue</th>
<th>Ground Truth</th>
<th>MAF</th>
<th>MOSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>KISMI: Bas na Sahil bhai, meri firki kheech rahe ho na!! (Enough brother Sahil, are you teasing me?!) SAHIL: Nah, nah, kya hai ki, mere CD ki collection mein na, ye train ke awaaj vali CD nahi hai... (No no, see I don’t have train’s sound in my CD collection...)</td>
<td>Sahil Kismi ko taunt maarta hai kyuki use rail gaadi ki awaaj sunni hai. (Sahil taunts Kismi that she wants to hear the sound of a train)</td>
<td>Sahil Kismi ko taunt maarta hai ki use pasand nahi. (Sahil taunts Kismi that he doesn’t like)</td>
<td>Sahil Kismi ko taunt maarta hai kyuki use rail gaadi ki awaaj sunni hai. (Sahil taunts Kismi that she wants to hear the sound of a train)</td>
</tr>
<tr>
<td>MONISHA: Say hello to Tommy the dog. (Say hello to Tommy the dog.) MAYA: Tumne iss kutte ka naam Tommy the dog rakha? (Did you name your dog Tommy the dog?)</td>
<td>Maya monisha ko tana marhti hai kyuki usne apne kutte ka naam tommy the dog rakha hai. (Maya says that her dog’s name is Tommy the dog.)</td>
<td>Maya kehti hai ki uske kutte ka naam tommy the dog rakha hai. (Maya says that her dog’s name is Tommy the dog.)</td>
<td>Maya taunts monisha kyuki usne apne kutte ka naam tommy the dog rakha hai. (Maya taunts Monisha that she has named her dog Tommy the dog.)</td>
</tr>
</tbody>
</table>

Table 5: Accuracy for the sarcasm source and target for BART-based systems.

<table>
<thead>
<tr>
<th></th>
<th>mBART</th>
<th>BART</th>
<th>MAF</th>
<th>MOSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>75</td>
<td>77.23</td>
<td>91.07</td>
<td>90.17</td>
</tr>
<tr>
<td>Target</td>
<td>45.33</td>
<td>52.67</td>
<td>46.42</td>
<td>56.69</td>
</tr>
</tbody>
</table>

Table 6: Human evaluation.

<table>
<thead>
<tr>
<th></th>
<th>Coherency</th>
<th>On topic</th>
<th>Capturing sarcasm</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBART</td>
<td>2.57</td>
<td>2.66</td>
<td>2.15</td>
</tr>
<tr>
<td>BART</td>
<td>2.73</td>
<td>2.56</td>
<td>2.18</td>
</tr>
<tr>
<td>MAF</td>
<td>3.03</td>
<td>3.11</td>
<td>2.77</td>
</tr>
<tr>
<td>MOSES</td>
<td>3.96</td>
<td>3.27</td>
<td>3.10</td>
</tr>
</tbody>
</table>

Explanations for correct structure and grammar. **On topic:** Measures the extent to which the generated explanation revolves around the dialogue topic. **Capturing sarcasm:** Estimates the level of emitted sarcasm being captured in the generated output.

We show the average score for the human evaluation parameters in Table 6. As illustrated, the proposed MOSES model exhibits more coherent, on topic, and sarcasm related explanations. However, there is still a scope for improvement, which can be taken up as future work.

**Understanding Affects with Explanation**

We study three understanding tasks in dialogues – sarcasm detection, humour identification, and emotion recognition using sWITS, hWITS, and eWITS, respectively. A trained SED system is used to obtain the explanations for all the instances present in these datasets. We show the qualitative analysis of the generated explanations by MOSES in the supplementary. To verify our hypothesis that sarcasm explanation helps affect understanding, we perform experiments with and without explanations, as explained in the subsequent sections.

**Sarcasm Detection.** We take a base RoBERTa model (Liu et al. 2019) and perform the task of sarcasm detection over sWITS. The experimentation is performed using three setups as described below:

1. When we do not provide any utterance explanation to the input dialogue.
2. When we provide utterances appended with their generated explanation at the training time. Plain dialogues are given at the testing time in this case.
3. When dialogue instances are appended with their corresponding explanations during train and test time.

---

4 The evaluators are fluent in English and their age ranges in 25-30 years.
Table 7: Experimental results on RoBERTa base when explanations generated by MOSES and MAF are used for completing the respective tasks. The first row indicates the performance without explanation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Use of Exp</th>
<th>Sarcasm</th>
<th>Humor</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Train</td>
<td>Test</td>
<td>P R F1 Acc</td>
</tr>
<tr>
<td>None</td>
<td>0</td>
<td>0.55  0.68  0.62  0.65 0.59  0.65 0.73  0.72 0.87</td>
<td>0.58 0.69  0.78  0.73 0.73  0.78 0.78  0.78 0.78</td>
<td></td>
</tr>
<tr>
<td>MAF</td>
<td>1</td>
<td>0.58  0.68  0.73  0.71 0.68  0.70 0.85  0.79 0.80</td>
<td>0.60 0.69  0.83  0.81 0.81  0.80 0.83  0.80 0.80</td>
<td></td>
</tr>
<tr>
<td>MOSES</td>
<td>1</td>
<td>0.65  0.71  0.78  0.76 0.73  0.77 0.83  0.81 0.80</td>
<td>0.70 0.83  0.76  0.73 0.77  0.75 0.83  0.81 0.80</td>
<td></td>
</tr>
</tbody>
</table>

Table 7 illustrates the results we obtain for all the settings for MOSES and the best baseline, MAF. As can be seen, RoBERTa obtains 92% F1 score when we do not use any explanations. However, with the use of the generated explanations by MOSES during the train time, we obtain an improvement of 6% F1-score. On the other hand, the best performance is achieved by the last case, where the input instances are appended with their corresponding explanations both at the train and test time, with an increase of 8% F1-score. Consistent to the results obtained by MOSES’s generation, MAF also reports an improved performance over no explanation model. However, the improvement shown by MAF is not at par with the improvement obtained by MOSES. These results directly support our hypothesis that utterance explanations can assist an efficient detection of sarcasm in the input instances.

**Humour Identification.** Another RoBERTa base is used to perform humour identification on hWITS. As for sarcasm detection, humour identification is also evaluated for the three setups described in the previous section. Table 7 illustrates the results obtained for the described setups. When no explanations are used during the training or testing time, we get an F1-score of 73%. This score is comparable to the performance we get when input instances are appended with their corresponding explanations generated by MOSES at the training time. This performance is boosted by 3% when the explanations are provided at the train/test time. However, it is important to note that the explanations generated by the MAF model resulted in a slightly decreased performance indicating the superiority of MOSES.

**Error Analysis**

**Quantitative.** To capture the improvement exhibited by explanations in affect understanding, we show the confusion matrices emitted by the understanding models with and without using explanations. Table 8 illustrates these matrices – and as can be seen, the methods with explanation obtains higher true positive rate with a decreased false positive and false negative rates for majority of the classes among sarcasm, humour, and emotion labels.

**Qualitative.** While quantitative results confirm that explanations assist in identifying affects efficiently, qualitative analysis can further corroborate this hypothesis. Table 9 shows one instance from the test set where the presence of explanation helps for all affective tasks in question. More such examples can be found in the supplementary.

**Conclusion**

The inability of existing systems to understand sarcasm results in a performance gap for various affect understanding tasks like emotion recognition, humour identification, and sarcasm detection. To mitigate this issue, we proposed MOSES to explore the task of Sarcasm Explanation in Dialogues (SED). MOSES takes multimodal code-mixed sarcastic conversation instances as input and results in a natural language explanation describing the sarcasm present in it. We further explored the effect of the generated explanations on three dialogue-based affect understanding tasks – sarcasm detection, humour identification, and emotion recognition. We observed that explanations improved the performance of all three tasks, thus verifying our hypothesis.
Acknowledgements
The authors acknowledge the support of the ihub-Anubhuti Foundation, set up under the NM-ICPS scheme of the DST, and CAI-IIITD.

References
Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 6144–6154.


