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

**Shivani Kumar [1], Ishani Mondal [2], Md Shad Akhtar [1], Tanmoy Chakraborty [3]**

[1]Indraprastha Institute of Information Technology Delhi, India
[2]University of Maryland, College Park
[3]Indian Institute of Technology Delhi, India
shivaniku@iiitd.ac.in, ishani340@gmail.com, shad.akhtar@iiitd.ac.in, tanchak@iitd.ac.in

**https://github.com/LCS2IIITD/MOSES.git**

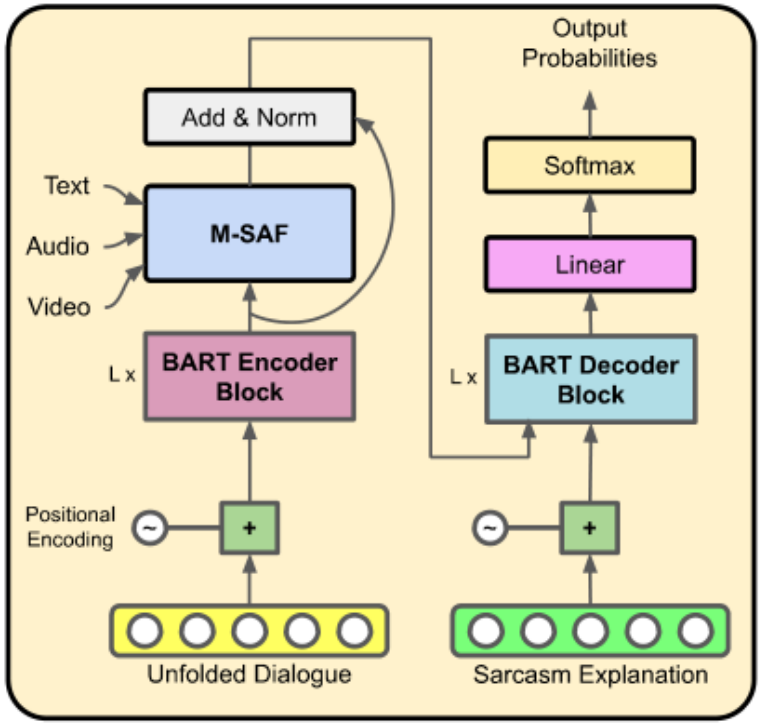
AAAI-2023

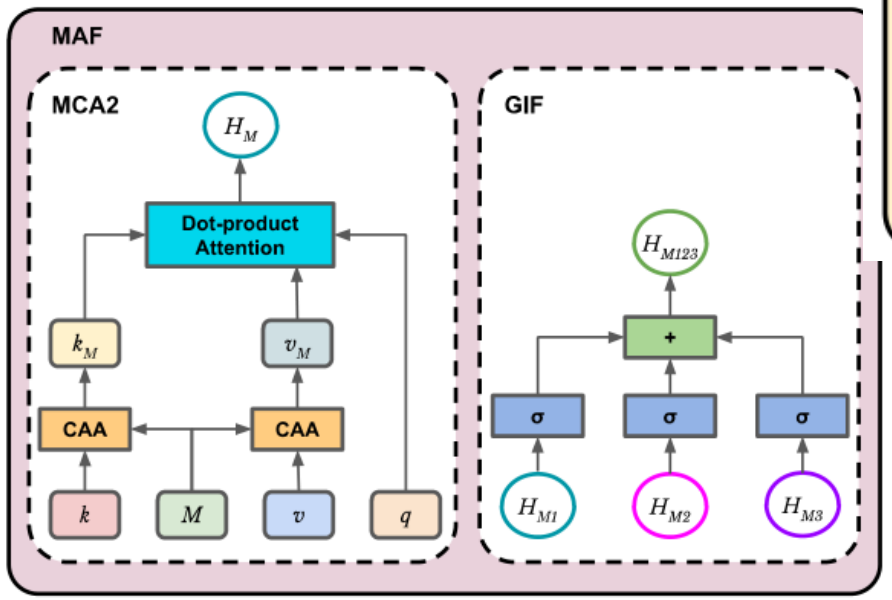
**Reported by JiaWei Cheng**

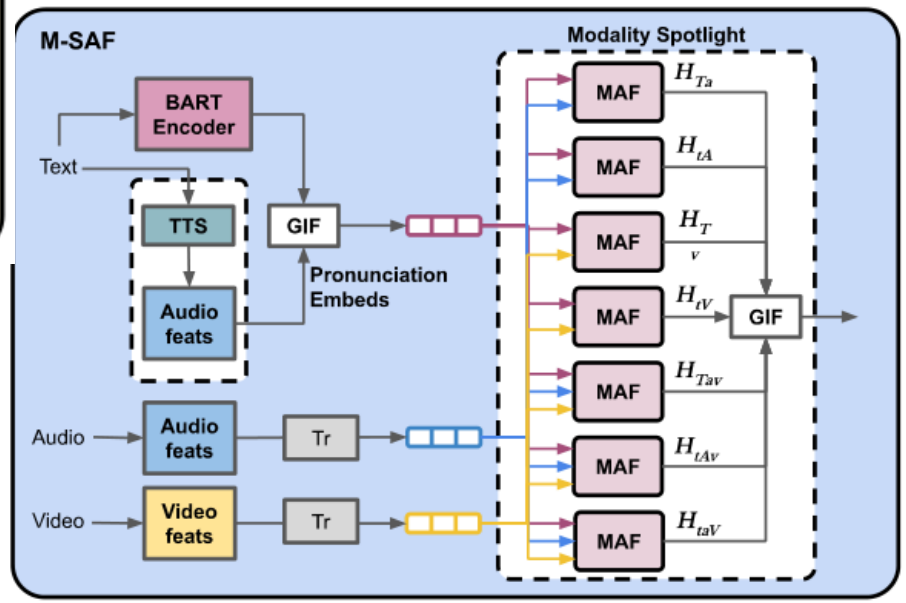
# Introduction


Figure 1: [Best viewed in color] Effect of multimodality on sarcasm. We do not show all possible combinations for brevity.

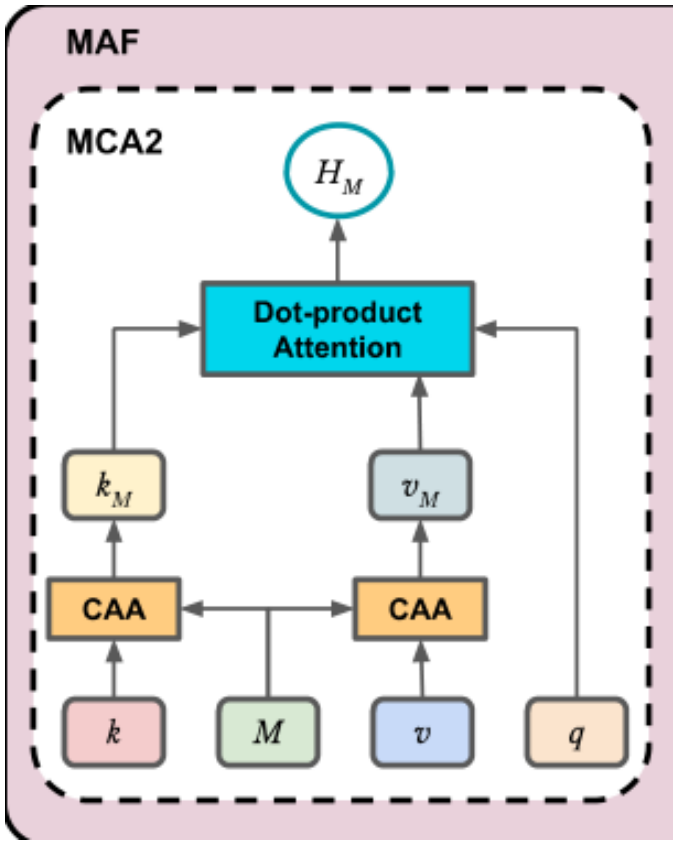
# Overview


Output Probabilities
Add & Norm
Softmax
Text
Audio
Video
M-SAF
Linear
L x
BART Encoder Block
L x
BART Decoder Block
Positional Encoding
Unfolded Dialogue
Sarcasm Explanation


MAF
MCA2
$H_M$
Dot-product Attention
$k_M$
$v_M$
CAA
CAA
$k$
$M$
$v$
$q$
GIF
$H_{M123}$
+
σ
σ
σ
$H_{M1}$
$H_{M2}$
$H_{M3}$


M-SAF
Modality Spotlight
BART Encoder
Text
TTS
Audio feats
GIF
Pronunciation Embeds
Audio
Audio feats
Tr
Video
Video feats
Tr
MAF
MAF
MAF
MAF
MAF
MAF
GIF
$H_{Ta}$
$H_{tA}$
$H_{T_v}$
$H_{tV}$
$H_{Tav}$
$H_{tAv}$
$H_{taV}$

# Method

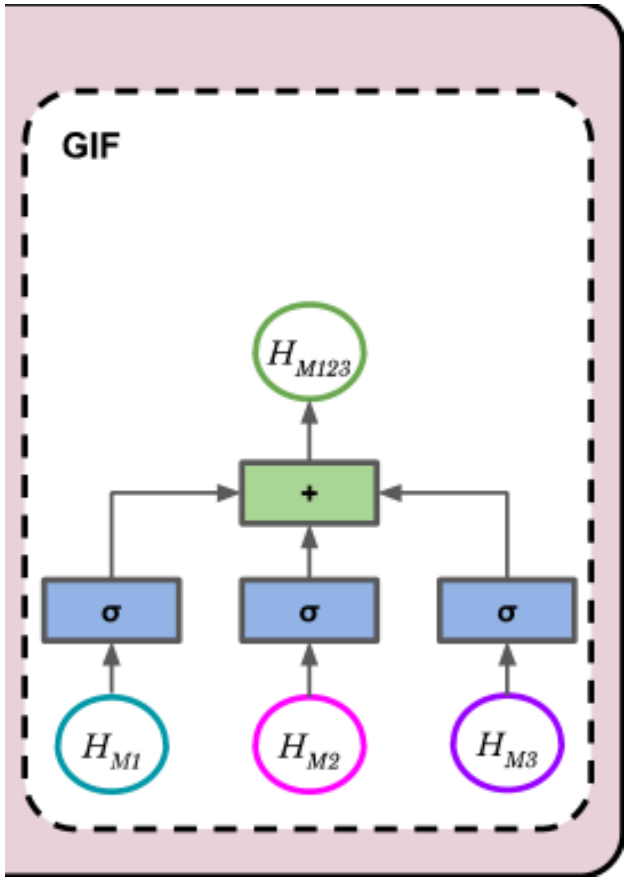



$$[qkv] = H[W_q W_k W_v] \tag{1}$$

$$\begin{bmatrix} k_m \\ v_m \end{bmatrix} = (1 - \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix}) \begin{bmatrix} k \\ v \end{bmatrix} + \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} (M \begin{bmatrix} U_k \\ U_v \end{bmatrix}) \tag{2}$$

$$\begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} = \sigma(\begin{bmatrix} k \\ v \end{bmatrix} \begin{bmatrix} W_{k_1} \\ W_{v_1} \end{bmatrix} + M \begin{bmatrix} U_k \\ U_v \end{bmatrix} \begin{bmatrix} W_{k_2} \\ W_{v_2} \end{bmatrix}) \tag{3}$$

# Method




$$g_a = [H \oplus H_a]W_a + b_a \tag{4}$$

$$H_{Ta} = H + g_a \odot H_a \tag{5}$$

$$\begin{aligned} g_a &= [H \oplus H_a]W_a + b_a \\ g_v &= [H \oplus H_v]W_v + b_v \\ H_{Tav} &= H + g_a \odot H_a + g_v \odot H_v \end{aligned}$$

$$\begin{aligned} H_{all} = g_t \odot H + g_{Ta} \odot H_{Ta} + g_{tA} \odot H_{tA} + \\ g_{Tv} \odot H_{Tv} + g_{tV} \odot H_{tV} + g_{Tav} \odot H_{Tav} + \\ g_{tAv} \odot H_{tAv} + g_{taV} \odot H_{taV} \end{aligned} \tag{6}$$

# Experiments

| Mode | Model | R1 | R2 | RL | B1 | B2 | B3 | B4 | M |
|---|---|---|---|---|---|---|---|---|---|
| Textual | RNN | 29.22 | 7.85 | 27.59 | 22.06 | 8.22 | 4.76 | 2.88 | 18.45 |
| | Transformer | 29.17 | 6.35 | 27.97 | 17.79 | 5.63 | 2.61 | 0.88 | 15.65 |
| | PGN | 23.37 | 4.83 | 17.46 | 17.32 | 6.68 | 1.58 | 0.52 | 23.54 |
| | mBART | 33.66 | 11.02 | 31.5 | 22.92 | 10.56 | 6.07 | 3.39 | 21.03 |
| | BART | 36.88 | 11.91 | 33.49 | 27.44 | 12.23 | 5.96 | 2.89 | 26.65 |
| Multimodality | MAF-TA | 38.21 | 14.53 | 35.97 | 30.58 | 15.36 | 9.63 | 5.96 | 27.71 |
| | MAF-TV | 37.48 | 15.38 | 35.64 | 30.28 | 16.89 | 10.33 | 6.55 | 28.24 |
| | MAF-TAV | 39.69 | 17.1 | 37.37 | 33.2 | 18.69 | 12.37 | 8.58 | 30.4 |
| | MOSES-TA | 38.27 | 14.53 | 35.72 | 31.57 | 16.37 | 9.66 | 6.06 | 29.27 |
| | MOSES-TV | 39.62 | 16.78 | 37.48 | 32.69 | 17.76 | 11.01 | 6.89 | 31.65 |
| | MOSES-TAV | 40.88 | 18.33 | 38.38 | 33.27 | 18.87 | 12.6 | 8.8 | 31.41 |
| | MOSES | **42.17** | **20.38** | **39.66** | **34.95** | **21.47** | **15.47** | **11.45** | **32.37** |

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.

# Experiments

| Model | R1 | R2 | RL | B1 | B2 | B3 | B4 | M |
|---|---|---|---|---|---|---|---|---|
| **BART** | 36.88 | 11.91 | 33.49 | 27.44 | 12.23 | 5.96 | 2.89 | 26.65 |
| **+concat** | 17.22 | 1.7 | 14.12 | 13.11 | 2.11 | 0.0 | 0.0 | 9.34 |
| **+DPA** | 36.43 | 13.04 | 33.75 | 28.73 | 14.02 | 8.0 | 4.89 | 25.6 |
| **+MCA2** | 36.37 | 13.85 | 34.92 | 28.49 | 14.34 | 9.0 | 6.16 | 25.75 |
| **+ GIF** | 39.69 | 17.1 | 37.37 | 33.2 | 18.69 | 12.37 | 8.58 | 30.4 |
| **+ PE** | 40.88 | 18.33 | 38.38 | 33.27 | 18.87 | 12.6 | 8.8 | 31.41 |
| **+ MS** (MOSES) | **42.17** | **20.38** | **39.66** | **34.95** | **21.47** | **15.47** | **11.45** | **32.37** |

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

# Experiments

| Dialogue | Ground Truth | MAF | MOSES |
|---|---|---|---|
| KISMI: Bas na Sahil bhai, meri firki kheech rahe ho na!? *(Enough brother Sahil, are you teasing me?!)* SAHIL: Nahi, nahi, 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...)* | 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)* | Sahil Kismi ko taunt maarta hai ki use pasand nahi. *(Sahil taunts Kismi that he doesn't like)* | 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)* |
| MADHUSUDHAN: Kitne saal ka ho jaaega vo? *(How old will he be?)* INDRAVARDHAN: Aap ko ka lena dena, panchaanyati laal! *(What does it have to do with you, Mr. Poke-a-nose?)* | Indravardhan Madhusudan ke questions se pareshaan hai. *(Indravardhan is irritated by Madhusudhan's questions)* | Indravardhan Madhusudan ke behare pan se pareshaan hai. *(Indravardhan is tired of Madhusudhan's deafness)* | Indravardhan Madhusudan se pareshaan hai. *(Indravardhan is tired of Madhusudhan)* |
| 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?)* | Maya monisha ko tana marti hai kyunki usne apne kutte ka naam tommy the dog rakha hai. *(Maya taunts Monisha on naming her dog Tommy the dog.)* | Maya kehti hai ki uske kutte ka naam tommy the dog rakha hai. *(Maya says that her dog's name is Tommy the dog.)* | Maya taunts monisha kyunki usne apne kutte ka naam tommy the dog rakha hai. *(Maya taunts Monisha that she has named her dog Tommy the dog.)* |

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

# Experiments

| | mBART | BART | MAF | MOSES |
|---|---|---|---|---|
| **Source** | 75 | 77.23 | **91.07** | 90.17 |
| **Target** | 45.33 | 52.67 | 46.42 | **56.69** |

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

| | Coherency | On topic | Capturing sarcasm |
|---|---|---|---|
| **mBART** | 2.57 | 2.66 | 2.15 |
| **BART** | 2.73 | 2.56 | 2.18 |
| **MAF** | 3.03 | 3.11 | 2.77 |
| **MOSES** | **3.96** | **3.27** | **3.10** |

Table 6: Human evaluation statistics – comparing different models.

# Experiments

| Model | Use of Explanation | | Sarcasm | | | | Humor | | | | Emotion | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | P | R | F1 | Acc | P | R | F1 | Acc | P | R | F1 |
| None | 0 | 0 | 0.57 | 0.68 | 0.62 | 0.57 | 0.69 | 0.78 | 0.73 | 0.87 | 0.8 | 0.78 | 0.78 |
| MAF | 1 | 0 | 0.58 | 0.73 | 0.65 | 0.6 | 0.57 | **0.87** | 0.69 | 0.81 | 0.78 | 0.78 | 0.78 |
| | 1 | 1 | 0.66 | 0.77 | 0.71 | 0.68 | 0.73 | 0.71 | 0.72 | 0.87 | 0.78 | **0.81** | 0.79 |
| MOSES | 1 | 0 | 0.65 | 0.71 | 0.68 | 0.66 | **0.84** | 0.63 | 0.72 | **0.89** | 0.79 | 0.78 | 0.78 |
| | 1 | 1 | **0.70** | **0.83** | **0.76** | **0.73** | 0.72 | 0.77 | **0.75** | 0.88 | **0.81** | 0.80 | **0.80** |

Table 7: Experimental results on RoBERTa base when explanations generated by MOSESand MAF are used for completing the respective tasks. The first row indicates the performance without explanation.

# Experiments

| | NS | S |
|---|---|---|
| **NS** | 137/100 | 81/117 |
| **S** | 39/70 | 185/153 |

(a) Sarcasm detection on sWITS.

| | NH | H |
|---|---|---|
| **NH** | 335/330 | 32/37 |
| **H** | 24/23 | 82/83 |

(b) Humour identification on hWITS.

| | Neutral | Sadness | Joy | Anger |
|---|---|---|---|---|
| **Neutral** | 148/137 | 13/23 | 18/19 | 16/16 |
| **Sadness** | 5/2 | 62/66 | 3/2 | 0/0 |
| **Joy** | 7/5 | 10/9 | 120/124 | 4/3 |
| **Anger** | 0/9 | 0/1 | 8/9 | 50/48 |

(c) Emotion recognition on eWITS.

Table 8: Confusion matrix of the systems with and without explanations.

# Experiments

| | | | |
|---|---|---|---|
| **Dialogue** | **MAYA:** And this time I thought lets have a theme party! *(And this time I thought lets have a theme party!)*<br>**MONISHA:** Animals! Hum log sab animals banenge! *(Animals! Let us all be animals this time!)*<br>**MAYA:** Mai hiran, Sahil horse, and Monisha chhipakalee! *(I'll be a deer, Sahil a horse, and Monisha can be a lizard!)* | | |
| **Exp** | Maya Monisha ko animal keh ke taunt maarti hai. *(Maya taunts Monisha by calling her an animal)* | | |
| | **Sarcasm** | **Humour** | **Emotion** |
| **GT** | 1 | 0 | Anger |
| **w/o Exp** | 0 | 1 | Neutral |
| **w Exp** | 1 | 0 | Anger |

Table 9: True and predicted labels for the three affect tasks with and without using MOSES's explanation.

# Thanks !