Exercise? I thought you said 'Extra Fries': Leveraging Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

Shraman Pramanick, Md. Shad Akhtar, Tanmoy Chakraborty
Laboratory for Computational Social Systems (LCS2)
Indraprastha Institute of Information Technology, New Delhi, India

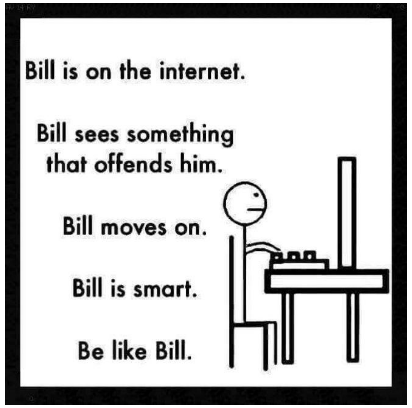
Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

- In recent years, Internet memes (or simply memes) have emerged as one of the most frequently circulated entities on social media platforms.

- Interpreting memes is a challenging task:
  - The semantics of memes often depend upon **implicit world knowledge**
  - Two memes can have **same image** (and vice versa) but can convey entirely **different semantics**
  - Annotating memes is challenging - "**Subjective Perception Problem**[1]"

Humor, Sarcasm, (-) Sentiment

Neutral



[1]Zhao, S.; Ding, G.; Huang, Q.; Chua, T.-S.; Schuller, B. W.; and Keutzer, K. 2018. Affective Image Content Analysis: A Comprehensive Survey. In IJCAI, 5534–5541

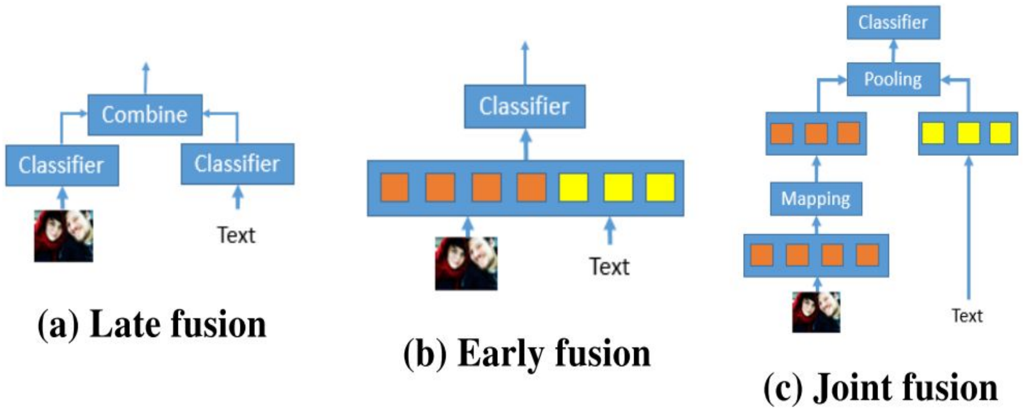
Problem Statement:

- A meme M is an image consisting of two modalities – a background image I and some text T at the foreground, referring to a specific situation.
- GIven a meme M, we analyze the emotion of memes on three dimensions:
  - **Sentiment classification** - positive, negative, or neutral
  - **Affect classification** - humor, sarcasm, offense, motivation, or a combination of the four affects
  - **Affect quantification** - what is the quantification of the expressed affect. {0, 1, 2, 3}
- **Dataset:** Memotion 1.0 dataset[2], released in SemEval-2020 shared task on 'Memotion Analysis'

[2]Sharma, C.; 2020. SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor! In SemEval-2020, 759–773.

Related Work:

- **Multimodal Fusion:** Early Fusion, Late Fusion, Hybrid Fusion
  - **Early Fusion** directly integrates multiple sources of data into a single feature vector
  - **Late Fusion** refers to the aggregation of decisions from multiple sentiment classifiers
  - **Hybrid Fusion** employs an intermediate shared representation



Visualization of fusion techniques (source: Duong et al.[3], 2017)

- **Meme Emotion Analysis: SemEval-2020 Task 8 Memotion Analysis** - top participants used FFNN, Naive Bayes, ELMo, MMBT, BERT for textual modality and Inception-ResNet, Polynet, DenseNet and PNASNet for visual modality.

[3]Duong C T.; 2017. Multimodal Classification for Analysing Social Media.

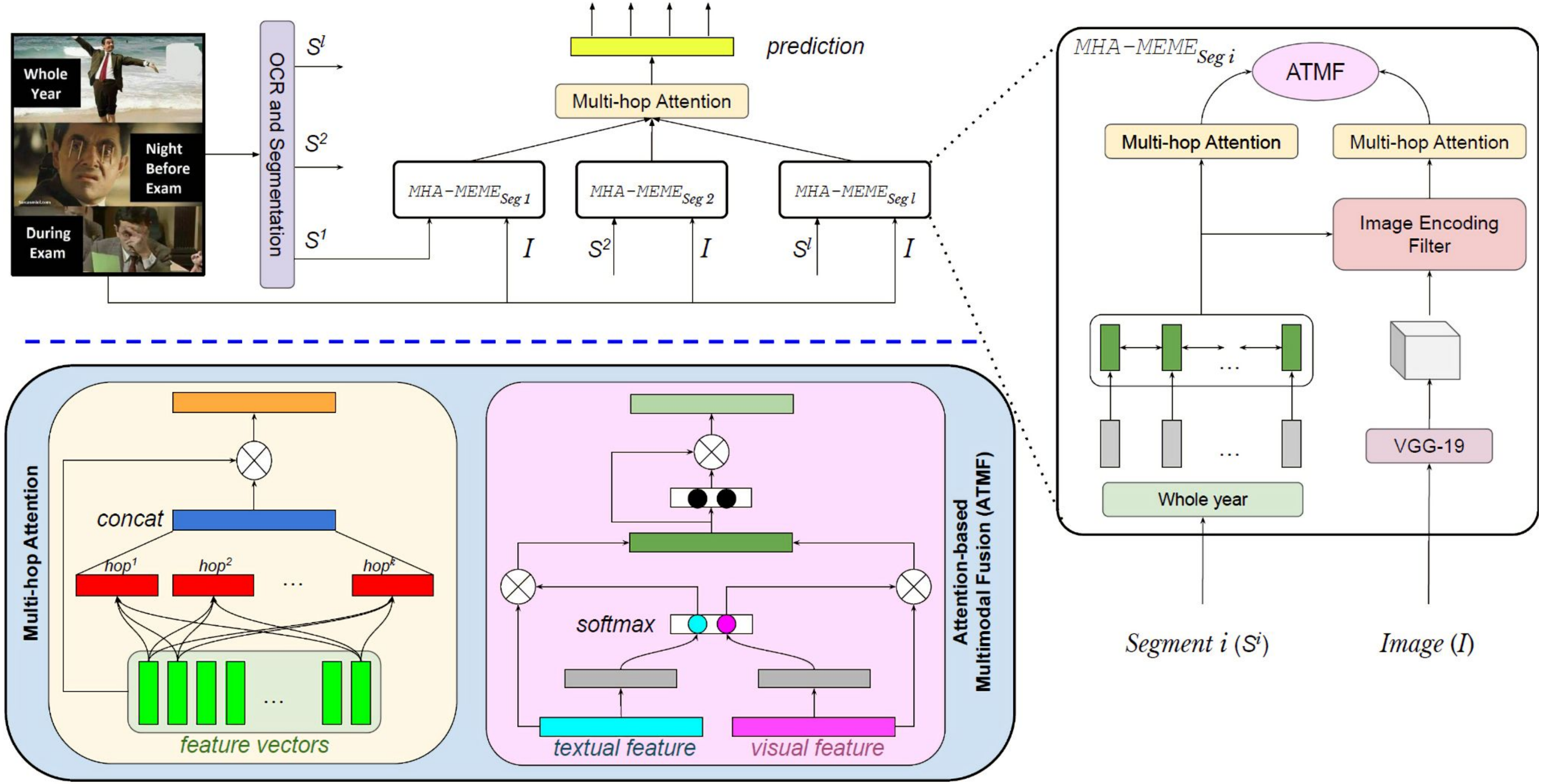
Contributions:

- We leverage correspondence between a meme and its constituent texts depending upon the spatial locations.
- We propose MHA-Meme, an attentive framework that effectively selects and utilizes complementary features from textual and visual modalities to capture multiple aspects of emotions expressed by a meme.
- We report state-of-the-art results for all the three tasks.
- We establish the interpretability of MHA-Meme using LIME framework.

ICWSM-2021
June 7-10, 2021

Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

MHA-Meme:



Principle Components:

- Text Encoder
- Image Encoder
- Multi-hop Attention
- ATMF
- Classifier

Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

- **<u>Text Encoder:</u>** BiLSTM [$H = (h_1, h_2, \ldots\ldots, h_n)$]
- **<u>Image Encoder:</u>** VGG-19 [$F = (f_1, f_2, \ldots\ldots, f_m)$]

- **<u>Image Encoding Filter:</u>**
  - We want to extract complementary features from textual and visual modality
  - The OCR-extracted text and the text in the image do not establish a direct correspondence
  - This Image Encoding Filter block outputs refined image features, $U = (u_1, u_2, \ldots\ldots, u_m)$, filtering redundant information from two modalities

- **<u>Multi-hop Attention:</u>** Originally proposed by Lin et al. (2017)[4] - helps in capturing all different semantics expressed by a meme; applied on top of image and text features.

- **<u>Attention-based Multi-modal Fusion (ATMF):</u>** Same modality may have different contribution for different meme samples; computes modality specific attention score.

[4]Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou,B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding.

LCS2 — Laboratory for Computational Social Systems | ICWSM-2021 June 7-10, 2021

Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

Implementation Details:

- Memotion Analysis' dataset contains 6601 training samples and 1879 test samples. Additionally, we to validate the generalizability of MHA-Meme, we collected and annotated an additional set of 334 memes.
- To alleviate data imbalance, we applied larger weights to minority classes in cross-entropy loss.

Our Implementation is publicly available at https://github.com/LCS2-IIITD/MHA-MEME

Scan here:

Hyper-parameter	Notation	Value
hidden units of BiLSTM	u	256
#dim for Dense layers	-	$[256, 64, 8]$
Multi-hop Attention		
#hops (unimodal)	k	30
#hops (multimodal)		10
#hidden-units (unimodal)	d	350
#hidden-units (multimodal)		100
Training		
Batch-size	-	8
Epochs	N	200
Optimizer	-	Adam
Loss	-	NLL
Learning-rate	α	0.005
Learning-rate-decay (/10K$iter$)	-	1e-4
Momentum	-	0.9
Class weights for imbalanced training data		
sentiment $[w_{pos}, w_{neu}, w_{neg}]$		$[1, 1.5, 2]$
affective - *humor* $[w_{nonhum}, w_{hum}]$		$[1.5, 1]$
affective - *sarcasm* $[w_{nonsar}, w_{sar}]$		$[1.5, 1]$
affective - *offense* $[w_{nonoff}, w_{off}]$		$[1.25, 1]$
affective - *motivation* $[w_{nonmot}, w_{mot}]$		$[1, 1.25]$

Table 2: Hyper-parameters of MHA-Meme.

Ablation Results:

	Models	Sentiment classification				Affect classication (Avg)				Affect quantification (Avg)			
		Macro F1		Micro F1		Macro F1		Micro F1		Macro F1		Micro F1	
		$Test_A$	$Test_B$	$Test_A$	$Test_B$	$Test_A$	$Test_B$	$Test_A$	$Test_B$	$Test_A$	$Test_B$	$Test_A$	$Test_B$
T	BiLSTM - OCR	0.338	0.373	0.509	0.572	0.421	0.455	0.542	0.570	0.302	0.310	0.420	0.438
	BERT - OCR	0.336	0.375	0.512	0.570	0.422	0.449	0.549	0.571	0.295	0.298	0.395	0.402
	BiLSTM - OCR_{Seg}	0.352	0.391	0.560	0.594	0.475	0.490	0.570	0.594	0.319	0.332	0.422	0.442
	BERT - OCR_{Seg}	0.351	0.384	0.538	0.580	0.471	0.482	0.563	0.581	0.311	0.316	0.418	0.425
I	InceptionV3	0.322	0.358	0.516	0.557	0.407	0.430	0.499	0.525	0.288	0.287	0.402	0.406
	V16	0.318	0.355	0.521	0.560	0.399	0.432	0.505	0.532	0.286	0.295	0.411	0.418
	V19	0.325	0.367	0.525	0.562	0.413	0.448	0.518	0.550	0.292	0.300	0.405	0.419
T+I	BERT - OCR_{Seg} + V19	0.356	0.410	0.585	0.624	0.508	0.529	0.620	0.645	0.325	**0.362**	0.424	0.435
	BiLSTM - OCR_{Seg} + V19	**0.376**	**0.426**	**0.608**	**0.635**	**0.523**	**0.545**	**0.682**	**0.698**	**0.333**	0.360	**0.430**	**0.444**

Table 3: Ablation results on multimodal inputs and various feature extraction mechanisms. For the affect classification and quantification tasks, we report average scores. T: Text, I: Image, V16: VGG16, V19: VGG19.

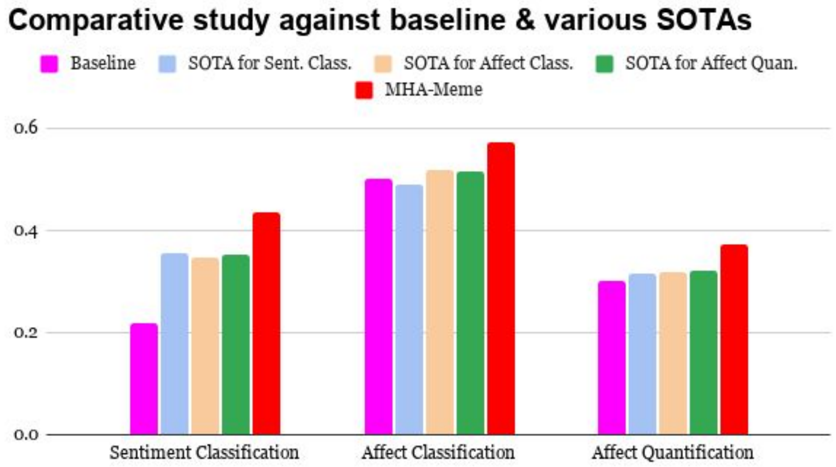
- Among unimodal systems, textual modality performs better than visual modality.
- Segmenting the OCR text helps in improving results.
- Multimodal systems outperform unimodal systems.

Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

State-of-the-art Results for Three Tasks:

System	Sent.	Affect classification					Affect quantification				
		Hum	Sar	Off	Motiv	Avg	Hum	Sar	Off	Motiv	Avg
Bs*	0.218	0.512	0.506	0.491	0.491	0.500	0.248	0.241	0.230	0.484	0.301
A*	0.352^2	0.515	0.511^3	0.512	0.520^3	0.515^2	**$0.271^{1\ddagger}$**	0.250	0.258	0.512	**$0.322^{1\ddagger}$**
B*	**$0.355^{1\ddagger}$**	0.473	0.508	0.499	0.474	0.489	0.262	**$0.259^{1\ddagger}$**	0.264^2	0.474	0.314
C*	0.345	0.516^3	**$0.516^{1\ddagger}$**	**$0.522^{2\ddagger}$**	0.519	**$0.518^{1\ddagger}$**	0.249	0.254	0.247	0.519	0.317^3
D*	0.341	0.521^2	0.441	0.491	0.512	0.491	0.264^3	0.254	0.241	0.517	0.319^2
E*	0.346	0.514	0.504	0.512	0.507	0.511^3	0.0	0.0	0.0	0.507	0.127
K.1*	0.350^3	-	-	-	-	-	-	-	-	-	-
F*	0.325	**$0.529^{1\dagger}$**	0.485	**$0.529^{1\dagger}$**	0.491	0.509	0.261	0.236	**$0.265^{1\ddagger}$**	0.491	0.313
G*	0.339	0.502	0.499	0.479	0.498	0.494	0.236	0.230	0.262^3	0.521^3	0.312
H*	0.323	0.493	0.487	0.505	0.490	0.494	0.237	0.255^2	0.252	0.502	0.311
I*	0.335	0.510	0.513^2	0.506	0.509	0.509	0.256	0.244	0.248	0.509	0.314
J*	0.345	0.434	0.447	0.400	0.488	0.442	0.255	0.254^3	0.241	0.488	0.310
K.2*	0.248	0.502	0.494	0.496	**$0.534^{1\dagger}$**	0.506	0.140	0.233	0.261	**$0.534^{1\dagger}$**	0.292
K.3*	0.323	0.486	0.500	0.472	0.522^2	0.495	0.215	0.193	0.233	0.522^2	0.291
K.4*	0.349	0.514	0.495	0.486	0.494	0.497	0.265^2	0.245	0.246	0.494	0.312
K.5*	0.337	0.500	0.483	0.516^3	0.520	0.505	0.251	0.238	0.256	0.520	0.316
MM	**$0.376^\dagger$**	**$0.527^\ddagger$**	**$0.520^\dagger$**	0.517	**$0.531^\ddagger$**	**$0.523^\dagger$**	**$0.271^\dagger$**	**$0.260^\dagger$**	**$0.268^\dagger$**	**$0.531^\ddagger$**	**$0.333^\dagger$**

Table 4: Comparative study against baselines and various state-of-the-art systems. All scores are Macro-F1 as per the official evaluation metric of the 'Memotion Analysis' shared task (Sharma et al. 2020). Superscripts [1],[2], and [3] denote official rank of the system in the shared task. For each case, the best and the second ranked scores among all systems are denoted by dagger(†) and double-dagger(‡), respectively. The first batch of results (after baseline, *Bs*) denotes a set of top three ranked systems for the three tasks (on average). System*: Values taken from Sharma et al. (2020). MM: `MHA-Meme`.



On average, MHA-MEME beats all the top performing systems in the SemEval-20 Memotion Analysis Challenge in a range of 1.5% - 3% Macro-F1 score

ICWSM-2021

June 7-10, 2021

Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

Importance of Different Modules:

	Hops	Sentiment classification						Affect classification						Affect quantification					
		D-Fusion		AT-Fusion		ATMF		D-Fusion		AT-Fusion		ATMF		D-Fusion		AT-Fusion		ATMF	
		T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B	T_A	T_B
M1	S	33.6	37.2	34.2	38.1	34.5	38.5	50.1	52.2	50.4	52.7	50.5	52.9	30.7	32.4	31.4	33.0	31.8	33.5
	M	34.0	37.5	34.4	38.6	34.9	38.9	50.3	52.6	50.5	53.0	50.8	53.2	31.5	33.1	31.9	33.8	32.0	34.3
M2	S	35.5	38.6	358	39.1	37.0	40.9	51.0	52.8	51.2	53.3	51.7	54.0	32.2	34.4	32.6	34.8	32.9	35.2
	M	36.4	40.5	37.2	41.3	**37.6**	**42.6**	51.3	53.0	51.4	53.6	**52.3**	**54.5**	32.4	34.6	32.7	35.1	**33.3**	**36.0**

Table 5: Comparative study of different fusion mechanisms and effect of single-hop attention vs multi-hop attention (in %). **M1:** BiLSTM - OCR + VGG19. **M2:** BiLSTM - OCR_{Seg} + VGG19. **S:** Single hop; **M:** Multi hops

- D-Fusion is direct concatenation; AT-Fusion is an attentive framework proposed by (Poria et al. 2017)[5]. Our proposed ATMF performs superior to D-Fusion and AT-Fusion in all experiments.
- The incorporation of multi-hops yields ~2% improvement for different model variants.

[5]Poria, S.; Cambria, E.;2017b. Multi-level multiple attentions for contextual multimodal sentiment analysis. In ICDM, 1033–1038. IEEE

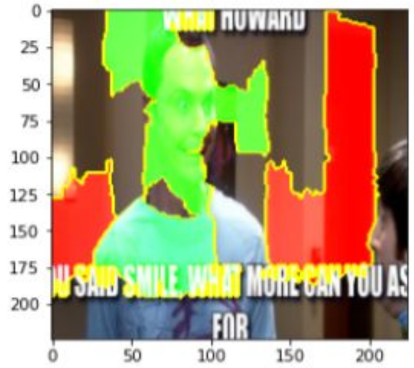
ICWSM-2021
June 7-10, 2021

Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis

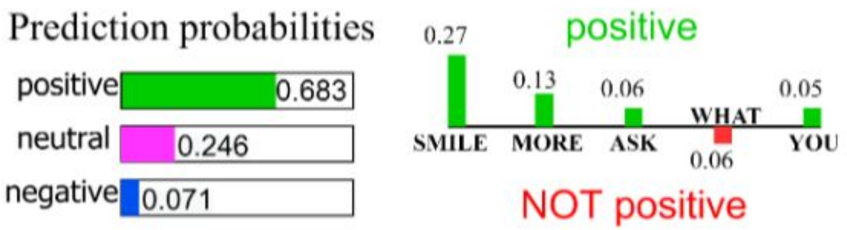
Interpretability of MHA-Meme:

- The prediction probabilities by MHA-Meme on this sample corresponding to positive, neutral, and negative sentiment classes are 0.683, 0.246, 0.071.
- The smiling face of the character, highlighted by green pixels, prominently contributes to the positive class.
- In text, the words ‘SMILE’ and ‘MORE’ imparts positive sentiment.
- The word ‘SMILE’ has highest attention weight in the two segments, supporting the explanations by the LIME framework.

(a) Input meme.



(b) LIME output - image.



(c) LIME output - text.

Segment$_1$: WHAT HOWARD
Segment$_2$: YOU SAID SMILE WHAT MORE CAN YOU ASK FOR

(d) Attention weights as computed by MHA−Meme.

Figure 3: Example of explanation by LIME on both visual and textual modalities and visualization of attention weights over text tokens obtained from MHA−Meme.

Conclusion:

- In this paper, we addressed three tasks related to the affect analysis of a meme, namely, sentiment classification, affect classification, and affect class quantification.
- We propose an attention-rich neural framework (called MHA-Meme) that analyzes the interaction between visual and textual modalities at fine-granular level. We design two attention mechanisms - a multi-hop attention module for the unimodal feature extraction and an attention-based multimodal fusion module for computing the interaction between the two modalities.
- MHA-Meme performs consistently across three tasks on Memotion Analysis dataset.
- In comparison, baseline systems did not report consistent performance for all the tasks or affect dimensions.

Thank You!