

DETECTING AND CHARACTERIZING HARMFUL MEMES: TECHNIQUES AND APPLICATIONS

SHIVAM SHARMA



**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2025**

© Indian Institute of Technology Delhi (IITD), New Delhi, 2025

Detecting and Characterizing Harmful Memes: Techniques and Applications

by

SHIVAM SHARMA

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

In fulfilment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

July 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Detecting and Characterizing Harmful Memes: Techniques and Applications**, submitted by **Shivam Sharma (2022EEZ8473)**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy (Ph.D.)**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Dr. Tanmoy Chakraborty

Associate Professor

Department of Electrical Engineering,

Indian Institute of Technology Delhi

Place: New Delhi

Date: July 1, 2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my PhD advisor, Dr. Tanmoy Chakraborty, whose guidance, encouragement, and vision orchestrated every step of this journey. I am also profoundly grateful to Dr. Shad Akhtar and Dr. Preslav Nakov for their invaluable insights and direction, which helped shape and refine my research. Additional shoutout is reserved for all the collaborators and contributors in the projects aligned with this Thesis. I am also grateful to Dr. Amitava Das, the entire Wipro R&D and leadership team, whose collaborative support created invaluable opportunities for me.

I extend my heartfelt appreciation to my institution's academic and non-academic staff, whose constant support has been instrumental in this journey. I sincerely appreciate my lab mates for their camaraderie and insightful discussions, which created a stimulating environment that greatly enhanced my work.

To my friends, though separated by miles, your steadfast support and encouragement never wavered; thank you for staying connected. I owe my strength to my parents and sisters—whose unwavering love has been my foundation. Finally, a special thank you to my wife and kid, who made this journey possible with their enduring patience, love and support. This accomplishment is as much yours as it is mine.

A handwritten signature in black ink, appearing to read 'Shivam', with a long horizontal stroke underneath.

Shivam Sharma

PhD Scholar

Department of Electrical Engineering,
Indian Institute of Technology Delhi

Place: New Delhi

Date: July 1, 2025

ABSTRACT

The rapid growth of online engagement and social media interactions has transformed communication dynamics, significantly increasing the volume and diversity of content generated across diverse social media platforms such as social networking sites, online group discussions, microblogging sites, image-centric platforms, and private messaging services. This surge in content has also heightened the risk of encountering harmful material, often presented in multimodal forms, including images, text, and videos. Among these, memes—particularly those combining text and images—have emerged as a powerful medium for conveying intricate and often subtle messages. While substantial research has focused on detecting hate speech on social media, addressing the computational challenges associated with multimodal and contextually nuanced content, such as *harmful memes*¹, remains a relatively underexplored area. This thesis explores the multifaceted aspects of analyzing memes within social media, emphasizing their *characterization, narrative role labelling, contextualization, and representation learning*. The study begins by characterizing harmful memes, identifying their satirical nature and the entities they target in a harmful manner, using novel multimodal models for detection and analysis. We then delve into the narrative roles within memes, focusing on identifying entities portrayed as victims, villains, or heroes or referred to in a harmful connotation in general, using a robust multimodal framework that models contextual saliency. Further, the thesis presents novel approaches for contextualizing memes, proposing tasks such as generating natural language explanations for visual narrative roles, dynamic context mining for augmenting meme understanding and question-answer-based formulation for memetic contextualization. Finally, the work advances representation learning by introducing self-supervised and affect-enriched pre-training methods that cater to the complexity of complex multimodal data, demonstrating the efficacy of these methods in improving meme analysis across various downstream tasks. Overall, the research addresses critical gaps in current methodologies and highlights the necessity of advanced multimodal approaches in understanding the nuanced interplay of text and imagery in memes.

¹**Note: Potentially sensitive content included; viewer discretion is requested.**

सार

ऑनलाइन जुड़ाव और सोशल मीडिया इंटरैक्शन की तेज़ी से बढ़ती प्रवृत्ति ने संचार की गतिशीलता को बदल दिया है, जिससे विभिन्न सोशल मीडिया प्लेटफॉर्म जैसे सोशल नेटवर्किंग साइट्स, ऑनलाइन समूह चर्चाएँ, माइक्रोब्लॉगिंग साइट्स, छवि-केंद्रित प्लेटफॉर्म और निजी मैसेजिंग सेवाओं पर निर्मित सामग्री की मात्रा और विविधता में उल्लेखनीय वृद्धि हुई है। इस सामग्री में वृद्धि ने हानिकारक सामग्री से सामना होने के जोखिम को भी बढ़ा दिया है, जो अक्सर बहुरूपीय रूपों में प्रस्तुत की जाती है, जिसमें छवियाँ, पाठ, और वीडियो शामिल होते हैं। इनमें से, मीम्स—विशेष रूप से जो पाठ और छवियों का संयोजन करते हैं—जटिल और अक्सर सूक्ष्म संदेशों को व्यक्त करने के लिए एक प्रभावी माध्यम के रूप में उभरे हैं। जबकि सोशल मीडिया पर हेट स्पीच का पता लगाने पर पर्याप्त शोध हुआ है, बहुरूपीय और संदर्भ-समृद्ध सामग्री, जैसे हानिकारक मीम्स¹, से संबंधित कम्प्यूटेशनल चुनौतियों को संबोधित करना अभी भी अपेक्षाकृत कम अध्ययनित क्षेत्र है। यह शोध प्रबंध सोशल मीडिया में मीम्स के विश्लेषण के विविध पहलुओं का अन्वेषण करता है, विशेष रूप से उनके चरित्रांकन, सेमांटिक भूमिका लेबलिंग, संदर्भीकरण, और प्रतिनिधित्व अधिगम पर बल देता है। यह अध्ययन हानिकारक मीम्स को विशेषता प्रदान करके शुरू होता है, उनके व्यंग्यात्मक स्वभाव और हानिकारक तरीके से लक्षित इकाइयों की पहचान करता है, और उनका पता लगाने और विश्लेषण के लिए नए बहुरूपीय मॉडल का उपयोग करता है। इसके बाद हम मीम्स में सेमांटिक भूमिकाओं का विश्लेषण करते हैं, उन इकाइयों की पहचान पर ध्यान केंद्रित करते हैं जिन्हें पीड़ित, खलनायक, या नायक के रूप में दर्शाया गया है या सामान्य रूप से हानिकारक संदर्भ में संदर्भित किया गया है, इसके लिए एक मजबूत बहुरूपीय ढांचा का उपयोग करते हुए जो संदर्भीय महत्ता को मॉडल करता है। इसके अतिरिक्त, यह शोध प्रबंध मीम्स के संदर्भीकरण के लिए नवीन दृष्टिकोण प्रस्तुत करता है, जैसे कि विज़ुअल सेमांटिक भूमिकाओं के लिए प्राकृतिक भाषा स्पष्टीकरण उत्पन्न करना, मीम समझ को संवर्धित करने के लिए डायनेमिक संदर्भ खनन, और मीमेटिक संदर्भीकरण के लिए प्रश्न-उत्तर आधारित सूत्रण। अंत में, यह कार्य प्रतिनिधित्व अधिगम को आगे बढ़ाता है, जटिल बहुरूपीय डेटा की जटिलता के लिए आत्म-पर्यवेक्षित और भावनात्मक-संपन्न प्री-ट्रेनिंग विधियों को प्रस्तुत करता है, जो विभिन्न डाउनस्ट्रीम कार्यों में मीम विश्लेषण में इन विधियों की प्रभावकारिता को प्रदर्शित करता है। कुल मिलाकर, यह शोध वर्तमान कार्यप्रणालियों में महत्वपूर्ण अंतराल को संबोधित करता है और मीम्स में पाठ और छवि के सूक्ष्म अंतःक्रिया को समझने में उन्नत बहुरूपीय दृष्टिकोणों की आवश्यकता पर प्रकाश डालता है।

¹ नोट: संभावित संवेदनशील सामग्री शामिल है; दर्शक विवेक का उपयोग करें।

TABLE OF CONTENTS

THESIS CERTIFICATE	i
ACKNOWLEDGEMENTS	ii
ABSTRACT	iii
LIST OF FIGURES	xi
LIST OF TABLES	xviii
ABBREVIATIONS	xxiii
NOTATIONS	xxv
1 Introduction	1
1.1 Understanding Harmful Memes	2
1.2 Motivation, Objective and Scope	4
1.3 Thesis Organization	10
2 Literature Review	11
2.1 Survey of Harmful Memes	11
2.2 Related Works	14
3 Characterizing Harmful Memes	17
3.1 Introduction	17
3.2 Dataset	19
3.2.1 Data Collection and Deduplication	19
3.2.2 Data Annotation	20

3.3	Proposed Methodology	21
3.4	Experiments	22
3.4.1	Baselines	22
3.5	Results	23
3.5.1	Harmful Meme Detection	24
3.5.2	Target Identification	24
3.5.3	Human Evaluation	25
3.5.4	Ablation Study	25
3.5.5	Transferability Assessment	26
3.6	Analysis	26
3.6.1	Side-by-side Diagnostics	27
3.6.2	Error Analysis	28
3.7	Conclusion and Future Work	29
4	Analyzing Narrative Framing of Entities	30
4.1	Introduction	30
4.2	Detecting Harmful Referencing in Memes	33
4.2.1	Dataset	34
4.2.2	Proposed Methodology	36
4.2.3	Experiments	37
4.2.4	Results	38
4.2.5	Analysis	41
4.3	Characterizing Entity Roles in Harmful Memes	43
4.3.1	Dataset	43
4.3.2	Proposed Methodology	44
4.3.3	Experiments	45
4.3.4	Results	47
4.3.5	Analysis	49

4.4	Conclusion and Future Work	51
5	Contextualizing Memes	53
5.1	Introduction	53
5.2	Explaining Visual Narrative Roles	57
5.2.1	Dataset	57
5.2.2	Proposed Methodology	58
5.2.3	Experiments	60
5.2.4	Results	61
5.2.5	Analysis	62
5.2.6	System Demo	64
5.3	Knowledge-Enriched Evidence Detection	67
5.3.1	Dataset	67
5.3.2	Proposed Methodology	68
5.3.3	Experiments	70
5.3.4	Results	71
5.3.5	Analysis	73
5.4	Rationale-Based Multimodal QA	75
5.4.1	Dataset	76
5.4.2	Proposed Methodology	77
5.4.3	Experiments	80
5.4.4	Results	80
5.4.5	Analysis	82
5.5	Conclusion and Future Work	85
6	Representation Learning for Meme Analysis	87
6.1	Introduction	87
6.2	Self-supervised Pre-training for Label-Efficient Analysis	89
6.2.1	Dataset	89

TABLE OF CONTENTS	viii
6.2.2 Proposed Methodology	90
6.2.3 Experiments	92
6.2.4 Results	93
6.2.5 Analysis	95
6.3 Affect-Enriched Multimodal Fusion for Emotion Detection	98
6.3.1 Dataset	98
6.3.2 Proposed Methodology	101
6.3.3 Experiments	103
6.3.4 Results	104
6.3.5 Analysis	106
6.4 Conclusion and Future Work	110
7 Ethics and Broader Impact	112
8 Conclusion and Future Scope	115
BIBLIOGRAPHY	116
Appendix	140
A Characterizing Harmful Memes	141
A.1 Implementation Details and Hyperparameter Values	141
A.2 Data Filtering	141
A.3 Annotation Guidelines	142
A.4 Lexical Statistics About HarMeme	146
B Analyzing Narrative Framing of Entities	148
B.1 Appendix for Detecting Harmful Referencing in Memes	148
B.1.1 Implementation Details and Hyper-parameter Values	148
B.1.2 Ablation Study	149
B.1.3 Annotation Guidelines	150

B.1.4	Ext-Harm-P Characteristics	152
B.2	Appendix for Characterizing Entity Roles in Harmful Memes	154
B.2.1	Implementation Details and Hyperparameter Values	154
B.2.2	Additional details of HVVMemes	155
B.2.3	Entity Role Prediction Analysis.	156
C	Contextualizing Memes	158
C.1	Appendix for Explaining Visual Narrative Roles: EXCLAIM	158
C.1.1	Implementation Details and Hyperparameter values	158
C.1.2	Discussion on NLG Evaluation Metrics	159
C.1.3	Additional details on ExHVV and Annotation	160
C.1.4	Analyzing Explanations for Image-only and Multimodal Models . .	162
C.2	Appendix for Knowledge-Enriched Evidence Detection: MEMEX	163
C.2.1	Implementation Details and Hyperparameter Values	163
C.2.2	MCC: Additional Details	164
C.2.3	Comparing contexts from KYM and MIME	169
C.2.4	Limitations	170
C.3	Appendix for Rationale-Based Multimodal QA: MemeMQA	171
C.3.1	Implementation Details and Hyperparameter Values	171
C.3.2	Motivation for ARSENAL's Design	173
C.3.3	Multimodal Analysis of ARSENAL	175
D	Representation Learning for Meme Analysis	178
D.1	Appendix for Self-supervised Pre-training for Label-efficient Analysis . . .	178
D.1.1	Implementation Details and Hyperparameter Values	178
D.1.2	Statistical Analysis of Datasets	179
D.1.3	Training Characteristics	180
D.2	Appendix for Affect-Enriched Multimodal Fusion for Emotion Detection . .	181
D.2.1	Implementation Details and Hyperparameter Values	181

D.2.2	Additional details of MOOD	181
D.2.3	ALFRED's Performance with 'Neutral' Memes	184
	LIST OF PUBLICATIONS	187
	BIO-DATA	190

LIST OF FIGURES

1.1	Examples from our HarMeme dataset. The labels are in the format [Intensity, Target]. For Intensity, {0, 1, 2} correspond to <i>harmless</i> , <i>partially harmful</i> , and <i>very harmful</i> , respectively. For Target, {~, 0, 1, 2, 3} correspond to harm targeted towards <i>no entity</i> , <i>individual</i> , <i>organization</i> , <i>community</i> , and <i>society</i> , respectively. Examples 1.1 (b) and (c) are harmful but neither hateful nor offensive (absence of clear visual or verbal cues, allowing for subjective interpretation). Example 1.1 (d) is harmful <i>and</i> offensive (clear insinuations through sarcasm). ²	2
1.2	Typology of harmful memes. We show in <i>grey color</i> the categories for which we found no memes and no research publications; dotted boxes indicate that this type of memes exist, but we found no publications trying to detect it.	3
1.3	Examples of harmful memes. (a) A meme that is <i>partially harmful</i> , but is arguably not so hateful or offensive. (b) A meme is subtly sarcastic overall, but when the image and the text are considered in isolation, it is offensive (possibly to animal and environment welfare agencies), making it <i>very harmful</i> when taken as a whole.	5
1.4	Examples of heroes , villains and victims , as portrayed within memes.	5
1.5	A novel task of EXCLAIM. Given a meme, a referred entity and its narrative role generate a natural language explanation that emulates human-like reasoning.	6
1.6	Example of non-trivial emotion expression in memes: (a,b) text modality is misleading; (c,d) same visuals, with emotion defined by the text.	6
1.7	A Mind Map, visually depicting the conceptual framework for the thesis. The figure captures the <i>objectives</i> , <i>datasets</i> , and key <i>highlights</i> across the contributory chapters of the thesis. Note: → indicates <i>repurposing</i> – source dataset → derived dataset.	9

3.1	Examples of harmful memes. (a) A meme that is <i>partially harmful</i> , but is arguably not so hateful or offensive. (b) A meme is subtly sarcastic overall, but when the image and the text are considered in isolation, it is offensive (possibly to animal and environment welfare agencies), making it <i>very harmful</i> when taken as a whole.	18
3.2	The architecture of our proposed model, MOMENTA.	21
3.3	Example of explanation by LIME on both visual and textual modalities and visualization of bias in V-BERT for both tasks.	26
3.4	Visualization of explanation generated by LIME on both modalities for MOMENTA and ViLBERT.	28
3.5	Misclassified example and explanation.	28
4.1	Examples of harmful/not-harmful targeting in memes. (a) A meme that targets Justin Trudeau in a <i>harmful</i> way, with a communal angle. (b) A <i>non-harmful</i> mention of Justin Trudeau as a benign humor.	31
4.2	Examples of heroes , villains and victims , as portrayed within memes.	32
4.3	Example meme, along with the Candidate , Harmful and Not-harmful entities.	34
4.4	Comparison plots for the top-5 <i>harmfully referenced</i> entities, for their harmful/non-harmful referencing in our dataset.	35
4.5	The architecture of our proposed approach DISARM. Here, $\mathbf{c}_{mm}$ is the multimodal representation used for the final classification.	35
4.6	Comparison of the attention-maps for DISARM [(a), (b) & (c)] and ViLBERT [(d)] using BertViz and Grad-CAM.	41
4.7	Comparison of attention maps for misclassification between DISARM [(a), (b) & (c)] and ViLBERT [(d)] using BertViz and Grad-CAM.	42
4.8	Schematic diagram of VECTOR. The <i>vision module</i> (left) consists of a ViT-based image encoder, and <i>commonsense module</i> (right) leverages GCN-based commonsense knowledge graph embedding. <i>Information fusion module</i> (top) fuses corresponding outputs toward final classification. OTK: Optimal Transport-based Kernel.	44
4.9	Interpretability analysis for VECTOR; proposed model and VisualBERT; best multimodal baseline.	50
4.10	Error analysis for VECTOR and VisualBERT.	51

5.1	A novel task of EXCLAIM. Given a meme, a referred entity and its narrative role, generate a natural language explanation that emulates human-like reasoning.	54
5.2	MEMEX – given a meme and a relevant context, the aim is to identify the <i>evidence</i> in the context that can succinctly explain the background of the meme, depicted above via emboldened (ground-truth evidences) and highlighted (model predictions) excerpt.	55
5.3	The MememQA task: Given an input meme and multiple choices, identify the correct answer and justify.	55
5.4	An illustration of LUMEN’s architecture, constituting modules: (1) Visual Recognition, (2) Entity Semantics, (3) Explanation Generation, and (4) Role-Label Prediction.	59
5.5	Illustration of explanation generation quality for unimodal text (U-T), unimodal image (U-I), multimodal (MM), and LUMEN. Left: U-T explains better; Right: LUMEN explains better. Prediction scheme: Correct , Incorrect and Partially-correct	62
5.6	An example of rater A’s preference (bounded by the blue box) in favour of Explanation Option 1 (LUMEN) over Option 2 (T5-based text-only model).	64
5.7	EXCLAIM system, showcasing modular pipeline for meme processing.	65
5.8	EXCLAIM interface demonstrating caption generation, meme content extraction, and role-labelling of entities.	66
5.9	The architecture of our proposed model, MIME. We obtain external knowledge-enriched multimodal meme representation using Knowledge-enriched Meme Encoder (KME 1). We make use of a Meme-Aware Transformer (MAT 2) and a Meme-Aware LSTM layer (MA-LSTM 3) to incorporate the meme information while processing the context smoothly.	68
5.10	Evidence detection from MMBT (top) and MIME (bottom) for a sample meme. The emboldened sentences in blue indicate ground-truth evidence and the highlighted sentences indicate model prediction	74
5.11	Prediction errors from MIME on three <i>test-set</i> samples. The emboldened sentences in blue indicate ground-truth evidence and the highlighted sentences indicate model prediction	74
5.12	A schematic diagram showing question-answer construction process in MememQACorpus, using entity and role-label information from ExHVV.	76

5.13	Description of the prompting setup for free-form synthetic question generation using the LLM, Llama-2-7b-chat. The randomly chosen question option is highlighted in yellow.	77
5.14	Comparison of various prompt configurations examined. Bar color scheme – Green: unifiedqa-t5-base, Magenta: unifiedqa-t5-large, and Blue: t5-large.	78
5.15	A schematic diagram of ARSENAL for the MemeMQA task ($\oplus$: fusing the information via concatenation).	78
5.16	Comparison of ARSENAL’s output for a sample meme, with four baselines. The LLaVA-based rationale depicted is used for generating the explanation by ARSENAL. The font color scheme is as follows: correct , incorrect , and partially-correct	82
5.17	An example of the error type committed by ARSENAL (proposed approach) vs the correct inferencing by the MM-COT-Based approach.	84
6.1	Example of memes: (a,b) text modality is misleading; (c,d) same image but emotion being differentiated by the text.	88
6.2	Solution architectures of multimodal self-supervision for memes. MM-SimCLR: Multimodal SimCLR (left); Ext-PIE-Net: Extended Pie-Net (right).	90
6.3	Comparison between the proposed method and baselines on Memotion tasks. The X-axis signifies the incremental supervision during fine-tuning.	96
6.4	Training performance comparison for different label fractions [1 % – 10 % – 20 % – 50 %] for Ext-PIE-Net (top row) and MM-SimCLR (bottom row) on Memotion tasks. Dominant curves are <i>smoothed</i> depictions of the actual curves in the background.	97
6.5	Comparison b/w the proposed method and baselines on HarMeme tasks. The X-axis signifies the incremental supervision during fine-tuning.	97
6.6	Examples of the images added to <i>extend</i> AffectNet dataset, especially towards capturing the <i>non-human</i> subjects within the meme visuals like, (a) <i>fearful</i> ‘Spongebob’, (b) <i>joyous</i> cat and (c) <i>sad</i> puppy, depicted in subFigure (a), (b) and (c), respectively.	99
6.7	Examples depicting memes for six basic <i>Ekman</i> emotions from our dataset, MOOD.	99

6.8	Word clouds depicting the category-wise lexicon comprising the embedded texts for memes in the MOOD dataset.	100
6.9	Normalized histograms of meme text length per class.	101
6.10	ALFRED’s model architecture. GCA: Gated Cross Attention module, GMF: Gated Multimodal Fusion module, $\otimes$: Low-rank Bilinear Pooling, $\oplus$: Concatenation.	102
6.11	Depiction of Interpretability Analysis [subFigures (a) and (b) – ALFRED correctly predicts an <i>anger</i> meme], Error Analysis [subFigures (c) and (d) – ALFRED incorrectly predicts a <i>fear</i> meme as a <i>surprise</i> meme].	107
6.12	Thematic overlap analysis via similarity matrix, with <i>five</i> example cases of inter-emotion overlap highlighted. Each x/y-axes label represents <i>topic id</i> , followed by a set of <i>three</i> corresponding <i>representative key-words</i> (topicid_kw1_kw2_kw3).	108
6.13	Performance comparison for ALFRED and <i>previous best</i> on HarMeme (US Politics (P)/Covid (C); 2/3 class classification) and Dank Memes tasks, demonstrating ALFRED’s transferability.	109
A.1	Examples of memes that we filtered out as part of the annotation process and the reason for that filtering: (a) not in English, (b) blurry text, (c) cartoon, (d) text-only, and (e) image-only.	142
A.2	Examples of different types of harmful memes.	143
A.3	Snapshot of the PyBossa GUI we used.	144
A.4	Statistics about the distribution of sources and labels in the two datasets.	145
A.5	Normalized histograms of meme text length per class for the two datasets.	146
B.1	Examples of memes depicting different types (a)–(e) of <i>harmful</i> targeting. Note: Int. – International.	151
B.2	Distributions of the OCR’s length for the memes of top-5 harmful references: harmful (Blue) and non-harmful (Orange). The depiction is for Individual: (a) and (d); Organization: (b) and (e); and Community: (c) and (f).	152
B.3	Edge case	155
B.4	Entity distribution - US Politics.	155
B.5	Entity-role distribution - US Politics.	156

B.6	Entity role prediction summary for HVVMemes’s (combined) test set. X-axis: Top 20 entities referenced; Y-axis: True positive rate (Recall). Values at each bar’s centre of each portion depict corresponding total counts.	157
C.1	Word Clouds for the explanations from the two annotators.	160
C.2	Example memes illustrating the challenges encountered during annotations. Left: Multiple layers of reasoning involved; Top-right: Lighthearted mockery; Bottom-right: Perspective ambiguity.	161
C.3	Illustration of explanation generation quality, for unimodal text (U-T), unimodal image (U-I), multimodal (MM), and LUMEN. Left: U-I explains better; Right: Multimodal explains better. Prediction scheme: Correct , Incorrect and Partially-correct	161
C.4	Annotator agreement assessment. For the U.S. Politics meme example (left), Annotator A provides a biased explanation; Annotator B gives a fair one. Both Annotators A and B provide reasonable explanations for the COVID-19 meme example (right).	162
C.5	Examples of discarded meme types: (a) Text-only, (b) Code-mixed, (c) Image-only and (d) Cartoon.	165
C.6	Distribution of attributes in MCC. The total number of sentences in context passage range between 2 and 16 and the number of evidence sentences in context range between 1 and 10. The most common source of context is Wikipedia.	165
C.7	A Screenshot of the Annotation Tool. Abbr. details for various offsets captured: CB: Character beginning, CE: Character end, WB: Word beginning, WE: Word end.	166
C.8	Top-20 prominent topics representing themes of the memetic content in MCC.	167
C.9	Comparison b/w the quality of the OCR-extracted text via (a) Tesseract OCR, and (b) Google OCR.	172
D.1	Distributions of the text’s length. Blue: Hateful/Harmful; Orange: Not-hateful/harmful.	179
D.2	Training performance comparison (<i>unsmoothed</i>) for different label fractions [1 % – 10 % – 20 % – 50 %] for Ext-PIE-Net (top row) and MM-SimCLR (bottom row) on Memotion tasks.	180

D.3	A collection of meme examples featuring <i>human subjects, pop culture references, animated characters, and animals</i> from MOOD, with representative topics as <i>teacher, class, school, homework, test, kid, and exam</i>	183
D.4	Analyzing ALFRED's performance <i>with neutral</i> category. (a) and (b) Confusion Matrices and F1-score for ALFRED's two variants; (c) Quantifying Confusion: $FN_{E_i \rightarrow N} - TP_{E_i \rightarrow E_i}$: Difference between the <i>false-negatives</i> (FN) w.r.t <i>neutral</i> class (N) and <i>true-positives</i> (TP) for each Ekman emotion category (E_i).	184
D.5	Top-48 prominent topics representing themes of the <i>visually depicted content</i> in MOOD's memes.	185
D.6	Top-48 prominent topics representing themes of the <i>textually embedded content (OCR)</i> in MOOD's memes.	186

LIST OF TABLES

2.1	Categorization of Harmful Memes with Key Types and Detection Strategies Across Textual, Visual, and Multimodal Approaches.	12
2.2	A Summary of the experimental results for automatically detecting harmful memes. Y/N: positive and negative class labels; VH: Very harmful, PH: Partially-harmful, NH: Non-harmful; Tar. Ident.: Target Identification; PC: Protected category identification; PC. AT. : Protected category attack type; Off. Int.: Offense intensity prediction; Off: Offensive; Prop.: Propaganda; Prop. Tech.: Propaganda techniques, Prop. Tech.: (T): Text, Prop. Tech.: (S): text span; CB Inci.: Cyberbullying Incidents; CMAF: Cross-modal attention fusion. Cl.T: Classification task; B: Binary, M: Multi-class, ML: Multi-class and Multilabel; TM: TamilMemes, iNet: ImageNet, Fl: Flickr. The table is published as part of thesis work [SAA ⁺ 22].	13
3.1	Statistics about Harm-C and Harm-P. Harmful memes are also annotated with a target. Note – V/P/N H: Very/Partial/Not Harmful (Not Harmful and Harmless are used interchangeably in this work); Ind.: Individual, Org.: Organization, Comm.: Community, and Soc.: Society.	20
3.2	Performance on the two tasks. We merge <i>very harmful</i> and <i>partially harmful</i> . [†] This row shows the human performance on the test, and the last row shows the improvement of MOMENTA over the best baseline. Note – UM: Uni Modal, MM: Multi Modal.	23
3.3	Performance for target identification. [†] This row shows the human performance on the test set. Note – UM: Uni Modal, MM: Multi Modal.	24
3.4	Transferability of the two best-performing baselines and MOMENTA on Harm-C, on Harm-P, and on the combination thereof. The models are trained on the dataset in the row and tested on the one in the column. All scores are Macro F1. H-2 [†] is 2-class, H-3 [†] is 3-class harmful meme detection, and Tar* is target identification. Blue indicates the best transferable results for each task-dataset combination.	25
4.1	Dataset (Ext-Harm-P) Summary. Note – H: Harmful; NH: Not Harmful.	33

4.2	Performance comparison of unimodal and multimodal models vs. DISARM (and its variants) on Test Sets A and B.	39
4.3	Performance comparison of unimodal and multimodal models vs. DISARM (and its variants) on Test Set C.	40
4.4	Summary of statistics for # memes and entities referenced within them in HVVMemes [SSK ⁺ 22a]. The original train/val/test split ratio of 80:10:10 (%) for memes was preserved.	43
4.5	Benchmarking results (0's indicate no correct prediction). $\Delta_{(x-y)}$: Performance difference between models X and Y, # $\uparrow$ and # $\downarrow$: Absolute increment and decrement respectively.	45
4.6	Ablation study: Comparing VECTOR and its sub-modules over <i>best</i> model results. CAT: Fusion via concatenation, OTK: Fusion via Optimal Transport Kernel, KG: Knowledge Graph.	48
4.7	Comparison to the results on the CONSTRAINT@ACL-2022 shared-task on HVVMemes.	49
5.1	ExHVV summary: entity-specific sample counts w.r.t. different <i>domains</i> and <i>splits</i>	57
5.2	Prescribed guidelines for ExHVV annotation.	58
5.3	Performance comparison for EXCLAIM, using gen-eval metrics across unimodal, multimodal and LUMEN. The top two scores across the metrics are presented in bold; <i>higher</i> scores are better; † represents the second-best model. Mod.: Modality; U/M M: Uni/Multi Modal.	60
5.4	Ablation results w.r.t. LUMEN and its components, evaluated using gen-eval. Top scores across the metrics are presented in bold; ▲ indicates <i>higher</i> scores are better.	61
5.5	Evaluation using gen-eval _c , that measures generated text's closeness with ground-truth. Top two scores across the metrics are presented in bold; ▲ indicates <i>higher</i> scores are better; † represents the second best model. . . .	62
5.6	Error Evaluation using gen-eval _g . Top two scores across metrics are presented in bold; ▼ indicates <i>lower</i> score are better; † represents the second best model.	63
5.7	Prescribed guidelines for MCC's annotation.	67

5.8	Comparison of different approaches on MCC. The last row shows the absolute improvement of MIME over MMBT (the best baseline). E-F: Early Fusion and V-BERT: VisualBERT.	71
5.9	Component-wise evaluation: each component contributes to the performance of MIME, while removing them inhibits it. T: Transformer, L: LSTM, BiL: Bi-LSTM and MA: Meme-Aware.	73
5.10	Benchmarking results for MemeMQA, comparing the proposed approach vs unimodal and multimodal baselines. Table Footnotes: highest , <u>second-highest</u> , ● : MM-CoT (w Lecture) – Best Baseline, and ★ : ARSENAL (proposed approach). ARSENAL variants – (a). <i>w Entity-Specific</i> : Utilizes rationale conditioned upon the answer predicted by the first module; and (b). <i>w Generic</i> : Utilizes generic rationale.	80
5.11	Averages and Std. Dev. of ARSENAL’s performance measured across primary evaluation metrics, over five independent runs.	81
5.12	Robustness Analysis: (a) Question Diversification ($\mathbf{Q}_{div}$); (b) Confounder Setting (three scenarios).	82
5.13	Error rate comparison between ARSENAL, MM-CoT, unimodal (image and text), and multimodal baselines.	84
6.1	Comparison between the proposed SSL method and baselines on the Hateful Memes dataset. † represents SSL baseline and ★ is for the proposed approach.	93
6.2	Comparison of SSL+FT with previous best and baseline for Memotion tasks. † represents SSL baseline, and ★ is for the proposed approach and ‡ (Previous best): best scores for the corresponding tasks.	94
6.3	Comparison of SSL+FT with previous best and baseline for Harm-P (HarMeme) task.	95
6.4	Summary of MOOD & Extended AffectNet.	98
6.5	Prescribed guidelines for MOOD’s annotation.	100
6.6	Comparison of ALFRED and baselines on the MOOD dataset. The last row shows the improvement of ALFRED (*) over the <i>early-fusion</i> based model, designated as the best baseline (†). Class-wise accuracy for the six <i>Ekman</i> emotions in MOOD is also reported. Please note: Mod. – Modality.	104

6.7	Task-wise and class-wise performance (Macro-F1) of different approaches on the <code>Memotion</code> tasks. The last row shows the improvement of ALFRED (*) over the <i>previous best</i> results (†) reported. Please note: Mod. – Modality. .	105
6.8	Ablation study for ALFRED on MOOD; emotion features (EMO), GMF and GCA (via DCA: Dense Co-attention) modules.	106
A.1	Values of the hyperparameters for the models we experimented with for HarMeme task.	141
A.2	The top-5 most frequent words in each class for the two tasks and datasets. The TF-IDF score for each word is shown in parentheses.	146
B.1	Hyperparameters summary. [BS→Batch Size; LR→Learning Rate; V/T-Enc→Vision/Text-Encoder; vit→vit-base-patch16-224-in21k; bert:→bert-base-uncased; xlnet→xlnet-base-uncased]. . . .	148
B.2	Ablation results for DISARM and its variants for Test Sets A, B, and C. . .	149
B.3	The top-5 most frequently referenced entities in each harmfulness class and their target categories. The total frequency for each word is shown in parentheses.	152
B.4	Hyperparameters summary for Detecting Hero, Villain, and Victim from Memes. [BS→Batch Size; LR→Learning Rate; V/T-Enc→Vision/Text-Encoder; vit→vit-base-patch16-224-in21k; bert:→bert-base-uncased; xlnet→xlnet-base-uncased; resnet→resnet50].	154
C.1	Hyper-parameters for EXCLAIM: learning rate (LR), batch size (BS) and trainable parameter count (<i>#Params</i>) for each model.	158
C.2	Hyper-parameters and per-batch <i>inference</i> runtime for each model for MEMEX.	163
C.3	Comparison of the contextual insights obtained from KYM (knowyourmeme.com, <i>top</i>) and the one generated by MIME (<i>bottom</i>) for a sample meme. Text blacked-out (■) is for obscuring the user’s identity; Emboldened sentences in blue indicate ground-truth evidences and the highlighted sentences indicate model prediction	168
C.4	Hyper-parameter details for MemeMQA.	171

D.1	Hyperparameter values for the experiments on representation learning via self-supervised pre-training.	178
D.2	The top-5 most frequent words and their tf-idf scores in each class.	179
D.3	Hyperparameters of different models explored for emotion detection in memes.	181
D.4	Top 10 most frequent words in each emotion class. The TF-IDF score is in the parenthesis.	182

ABBREVIATIONS

HarMeme	Harmful Memes
Harm-C	Harmful Memes (Covid-19)
Harm-P	Harmful Memes (US Politics)
MOMENTA	Multimodal framework for detecting harmful MemEs aNd Their tArgets
Ext-Harm-P	Extended Harmful Memes (US Politics)
CE	Contextualized Entity
EH	Embedded Harmfulness
CI	Contextualized Image
CT	Contextualized Text
CMM	Contextualized Multimodal <i>embedding</i>
MMLRBP	Multimodal Low-rank Bilinear Pooling
DISARM	Detecting vIctimS targeted by hARmful Memes
HVVMemes	Hero-Villain-Victim Memes
VECTOR	Visual-narrative role dEteCToR
EXCLAIM	gENERating eXplanations for visual narrative roLe lAbeling In Memes
ExHVV	Explaining Hero-Villain-Victim Memes
LUMEN	expLaining visUal seMantic rolEs iN memes
gen-eval	evaluation suite for <i>benchmarking</i>
gen-eval_s	evaluation suite for <i>semantic-similarity-based</i> comparison of top-performing systems
gen-eval_e	evaluation suite for <i>error-rate-based</i> metrics of top-performing systems
MCC	Meme Context Corpus
MIME	MultImodal Meme Explainer
MEMEX	Task aimed to identify explanatory evidence for memes from their related contexts.
MemeMQA	Multimodal Question Answering for Memes
MM-SimCLR	Simple framework for multimodal contrastive learning of visual representations
Ext-PIE-Net	Extended Progressively Integrated Encoder Network
Memotion	Meme emotion analysis task
MMHS150K	Multimodal hate speech dataset
SSL+FT	Self-supervised learning, followed by fine-tuning
FT	Fine tuning
SSL	Self-supervised learning
SL	Supervised learning
SENT	Sentiment Analysis
EMOT	Emotion Detection

EMOT-Q	Emotion Quantification
MOOD	Meme emOtions Dataset
ALFRED	emotion-Aware muLtimodal Fusion foR Emotion Detection
LLM	Large Language Model
VLM	Vision Language Model

NOTATIONS

$\mathcal{L}$	Loss function
β_i	Weighing parameter # i
$\Delta(\%)$	% difference between the proposed model and the strongest baseline
$\uparrow x\%$	$x\%$ increment
$\downarrow x\%$	$x\%$ decrement
▲	The <i>higher</i> the measure, the better the performance
▼	The <i>lower</i> the measure, the better the performance