Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification

Abdullah Mazhar
IIIT Delhi
New Delhi, India
abdullahm@iiitd.ac.in

Zuhair Hasan Shaik
IIIT Dharwad
Dharwad, India
zuhashaik12@gmail.com

Aseem Srivastava
IIIT Delhi
New Delhi, India
aseems@iiitd.ac.in

Polly Ruhnke
University of Illinois Chicago
Chicago, USA
pruhn2@uic.edu

Lavanya Vaddavalli
University of Illinois Chicago
Chicago, USA
lvadd@uic.edu

Sri Keshav Katragadda
University of Illinois Chicago
Chicago, USA
skatra5@uic.edu

Shweta Yadav
University of Illinois Chicago
Chicago, USA
shwetay@uic.edu

Md Shad Akhtar
IIIT Delhi
New Delhi, India
shad.akhtar@iiitd.ac.in

Abstract

The expression of mental health symptoms through non-traditional means, such as memes, has gained remarkable attention over the past few years, with users often highlighting their mental health struggles through figurative intricacies within memes. While humans rely on commonsense knowledge to interpret these complex expressions, current Multimodal Language Models (MLMs) struggle to capture these figurative aspects inherent in memes. To address this gap, we introduce a novel dataset, AxiOM, derived from the GAD anxiety questionnaire, which categorizes memes into six fine-grained anxiety symptoms. Next, we propose a commonsense and domain-enriched framework, M3H, to enhance MLMs' ability to interpret figurative language and commonsense knowledge. The overarching goal remains to first understand and then classify the mental health symptoms expressed in memes. We benchmark M3H against 6 competitive baselines (with 20 variations), demonstrating improvements in both quantitative and qualitative metrics, including a detailed human evaluation. We observe a clear improvement of 4.20% and 4.66% on weighted-F1 metric. To assess the generalizability, we perform extensive experiments on a public dataset, RESTORE, for depressive symptom identification, presenting an extensive ablation study that highlights the contribution of each module in both datasets. Our findings reveal limitations in existing models and the advantage of employing commonsense to enhance figurative understanding.

CCS Concepts

• **Computing methodologies → Discourse, dialogue and pragmatics**; **Natural language generation**.

ACM Reference Format:
Abdullah Mazhar, Zuhair Hasan Shaik, Aseem Srivastava, Polly Ruhnke, Lavanya Vaddavalli, Sri Keshav Katragadda, Shweta Yadav, and Md Shad Akhtar. 2025. Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3696410.3714778

WWW '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714778

1 Introduction

In recent years, the increased use of social media platforms such as Reddit for sharing memes has emerged as a fascinating form of communication, especially among individuals expressing mental health symptoms [1]. These seemingly light-hearted memes often carry deep, hidden meanings that reflect complex emotional states. For many in contemporary society, memes are not just a source of humor but a creative outlet to subtly convey thoughts and feelings that are difficult to articulate directly. Yet, the true essence of these memes, especially those with mental health nuances, remains elusive to many. This happens primarily due to the inherent figurative language they often contain. To truly comprehend the core meaning behind these memes, humans require two major fronts: (a) commonsense knowledge and (b) figurative understanding. Together, these elements decode the deeper meaning embedded within a meme, revealing not only its humor but also the domain context behind it. For example, as shown in Figure 1, the person is sitting in a corner, holding head in distress, which suggests deep emotional pain or hopelessness. The text, *"I don't know how much more I can take"*, hints at overwhelming mental strain, possibly even suicidal thoughts. Here, commonsense understanding helps us interpret the person's body language and posture as signs of despair. Also, the figurative understanding is captured by the darkness of the image, symbolizing the person's inner turmoil and dark thoughts they may be experiencing. However, this process of interpretation, while relatively straightforward for human cognition, becomes significantly more complex when applied to artificial intelligence (AI).

Current advancements in visual language models (VLMs) and large language models (LLMs) continue to struggle with understanding sarcasm, satire, and nuanced figurative expressions in memes, especially in the context of mental health. These models have difficulty capturing the layered meanings that are essential for understanding such content [47]. Other works in this space have explored various facets of meme research. For instance, Wu et al. [42] developed

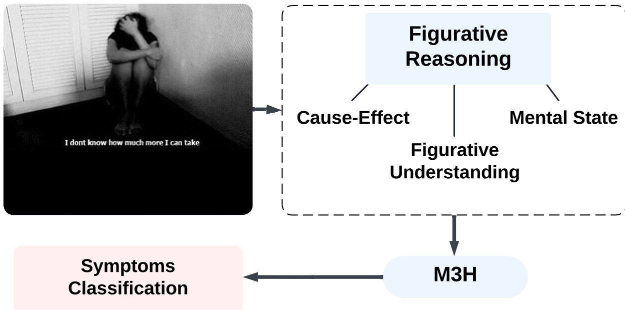


Figure 1: Our problem statement involves the classification of mental health memes focusing on depressive and anxious social media content. We employ figurative reasoning in our proposed method, M3H, for symptom classification.

a hateful meme detection method, focusing on the complexity of weakly related visual and textual modalities in memes. Other works show enhancement in meme detection by leveraging global and local perspectives and addressing harmful memes through multimodal analysis [32]. Sharma et al. [35] further surveyed the complexities of harmful meme detection, highlighting the gaps in datasets, particularly for under-explored topics like self-harm and extremism. However, we barely find research dedicated to the study of memes for sensitive domains like mental health. On the other hand, the primary cause of limited research in this space is possibly due to the limited datasets [24]. To the best of our knowledge, we find RESTORE to be the only publicly available dataset for depressive meme classification [44]. To address this gap, we first propose a novel dataset, AxiOM, containing memes derived from the GAD anxiety questionnaire. We choose this issue to introduce diversity in the existing set of data. AxiOM consists a total of 3582 memes. We employ annotators to categorize each meme into anxiety symptoms. Healthcare professionals further validate the efficacy of our annotation guidelines and the quality of the dataset produced. Next, we propose a novel multimodal framework, M3H, to classify mental health symptoms associated with each meme. We focus on employing figurative understanding with commonsense reasoning using VLMs. The core of our framework lies in utilizing these VLM's capabilities and further enhancing them through the dedicated retrieval-augmented-generation (RAG) method, allowing us to navigate the complex humor and underlying emotions within memes through knowledge fusion. For instance, as shown in Figure 1, our method first attempts to fetch figurative reasoning through commonsense with three major attributes: (a) cause-effect, (b) figurative-understanding, and (c) mental-state. Next, a classifier is responsible for symptom classification.

In our work, M3H contains GPT-4o for figurative reasoning and a BART-based transformer that allows the final encoding and classification of symptoms. We rigorously benchmark M3H against 6 competitive baseline methods (with 20 variations), including the current state-of-the-art system, on both the proposed AxiOM and RESTORE dataset. We employ standard classification metrics such as macro-F1 and weighted-F1 scores to evaluate performance, revealing significant improvements of 4.94% and 5.79% on macro-F1; whereas an improvement of 4.20% and 4.66% on weighted-F1 across AxiOM and RESTORE, respectively. Additionally, we present an exhaustive human evaluation and qualitative and quantitative analysis to further understand where our method excels and what it lacks. Our findings state the efficacy of our approach in bridging the gap between human-level meme comprehension and AI's current capabilities. Our contributions are summarized below:

- **Novel Dataset**: We propose a novel dataset, AxiOM, derived from the GAD anxiety questionnaire, which categorizes memes into six fine-grained anxiety symptoms.
- **Novel Framework**:We propose a novel framework, M3H, that employs figurative-cum-commonsense understanding for enhanced interpretation of memes related to mental health subjects and further classify associated symptoms.
- **Benchmarking**: We compare the performance of M3H against 6 baselines (with 20 variations) and provide extensive ablation studies and human evaluations to test the efficacy of our method.
- **Generalizability**: To validate the generalizability of our proposed model, we benchmark it on an additional dataset, RESTORE, where M3H consistently outperforms baseline methods.

Reproducibility. The code and dataset (for research purposes only) are available at https://github.com/flamenlp/M3H.

2 Related Work

Our work builds upon three key areas of prior research: the study of memes and social media, mental health expressions in online spaces, and multimodal approaches to understanding mental health content. Here, we review the most relevant studies in each of these domains.

Social Media and Memes. Memes have become a widely-used medium for expressing ideas and communicating on social media platforms. Researchers have developed techniques to detect and categorize memes within large datasets of social media content, leveraging the meme's visual and textual elements, metadata, and patterns of diffusion across networks [31, 36, 39]. Other studies have focused on the cross-platform spread of memes, providing valuable insights into how they are shared and reused across various social media environments [17]. The phenomenon of memes going "viral" has also been extensively examined, with researchers discovering that even simple images or phrases can rapidly gain traction online, forming temporary communities centered around their shared content [27, 34, 40]. In addition, some studies have specifically analyzed video memes, exploring how short video clips are repeatedly remixed and repurposed, often as a means of disseminating news [9, 20, 43]. Memes have also proven to be a valuable tool in marketing. Research indicates that humorous memes tend to be more effective than serious imagery in social media marketing campaigns, particularly when they garner high audience engagement [25, 41].

Social Media and Mental Health. The impact of social media on mental health, especially for young people, is an important area of study. One study found that when Facebook was introduced at colleges, it had a negative effect on students' mental health [10, 16]. However, other research on young adults has shown mixed results. Some studies suggest social media might harm mental health, while others find no evidence of harm or even some benefits [1, 11, 18, 29, 37]. Researchers have also looked at ways to predict mental health issues using social media data. However, a review of these methods found that there's a need for better standards and more valid ways of measuring mental health in these studies [3, 12, 33, 44]. During the COVID-19 pandemic, researchers found that people who used social media a lot were more likely to have mental health problems [11].

Multimodal Approaches and Commonsense Reasoning in Mental Health. Recent research has started to use information from multiple sources to understand mental health better. Previous studies used information from text, images, and how people interact on social media to analyze mental health [5, 7, 23]. Some researchers are using data from social media activities and physical signals (like heart rate) to monitor mental health without being intrusive [14, 48]. This approach could help detect mental health issues early. A new method has been developed to detect emotions and emotional reasoning in conversations [2]. It uses information from different sources and applies commonsense reasoning. This is particularly relevant to our work, as understanding memes often requires similar skills. Recently, researchers have proposed a new way to classify mental health using multiple types of data and a technique called "knowledge distillation" [19]. Additionally, Srivastava et al. [38] demonstrated the effectiveness of LLMs in mental health tasks like counseling summarization, further highlighting the potential of structured knowledge alignment. Our work builds on these studies by combining the analysis of memes, the focus on mental health in online spaces, and the use of multiple types of data and commonsense reasoning. We aim to create a more complete understanding of how depressive symptoms are expressed through memes.

Depression Analysis. The RESTORE dataset has been instrumental in advancing depression analysis through memes. This dataset is unique in its annotation of fine-grained depression symptoms based on the clinically adopted PHQ-9 questionnaire [26]. We perform sota work upon this dataset, and **our findings mention that we need to consider both visual and textual commonsense together at once. Traditional multimodal approaches that encode visual and textual information separately before fusing them may miss crucial implicit connections between these elements.**

3 Dataset

In this section, we discuss the dataset construction pipeline and the measures taken to ensure the quality of annotations followed by analysis. We gather data from various online sources, including Reddit and Instagram. We discuss the standard collection process and standards in Appendix (c.f Section A).

3.1 Annotation Guidelines

We utilize the *Labelbox* interface to annotate the memes based on the GAD questionnaire, which assesses anxiety severity. With the help of experts, we annotate the memes into six categories, each representing a symptom of anxiety. We establish the annotation guidelines in consultation with healthcare professionals and individuals diagnosed with anxiety. These definitions helped our annotators categorize memes according to the specific GAD symptoms. Below, we describe the definitions of each label.

- **Nervousness:** This category represents mild anxiety, typically related to common, short-term stressors. Memes in this category depict minor worries that cause nervousness but are manageable and not excessive.
  - Indications: This GAD category carries an indication of a person rubbing their forehead, temples, or hands. The text talks about something minor that made a person nervous but clearly won't last long, isn't excessive, and is manageable.
  - Example: A meme featuring a scared-looking man with the caption *"when you have to talk to the teacher about retaking a test to save your grade. This little maneuver is gonna cost us AN ANXIETY ATTACK"* showcases minor, temporary nervousness.
- **Lack of Worry Control:** This refers to anxiety that appears out of nowhere or feels intrusive, often overwhelming the individual. Memes in this category often reflect uncontrollable worry or the use of unhealthy coping mechanisms, such as substance use.
  - Indications: This sub-category is entailed by a visual of something or someone jumping into the picture or lurking in the background. The text is along the lines of *"My anxiety:"* to show the lack of control over their anxiety.
  - Example: A meme with the text *"Me: Wanting to add someone's phone number"* paired with an image of a worried character being stopped by an anxious figure saying *"What if you accidentally call them?"* implies a sudden, uncontrollable spike.
- **Excessive Worry:** This category involves persistent, excessive worry about both trivial and significant matters. Memes in this category typically feature visuals of profuse sweating or expressions of extreme stress.
  - Indications: This GAD category is entailed by the visuals having excessive sweating or shaking. The text expresses stress.
  - Example: A meme featuring a pink frog sitting on a lily pad with the text *"I spend most of my time worrying instead of doing"* and *"And that worries me"* reflects a sense of self-anxiety.
- **Difficulty Relaxing:** Memes in this category depict individuals struggling to rest or relax, often due to racing thoughts. These memes may feature visuals of someone unable to sleep or unwind despite being in a relaxing setting.
  - Indications: This GAD category may be entailed by the visuals of a person in a relaxing setting, like a bed or beach, with eyes wide. It is usually accompanied by text about overthinking, but because they can't seem to relax, it falls into the category of difficulty relaxing instead of excessive worry. Sometimes, the image can be of someone staring off into space with the text saying something along the lines of "Me at five AM:". This means they cannot rest and, therefore, are categorized as having difficulty relaxing.
  - Example: A meme featuring a four-panel meme showing a brain talking to a person in bed. In the first panel, the brain says, *"So you realized you suffer from anxiety?"* to which the person responds *"Yes"* while trying to sleep. In the next panel, the brain follows up with, *"But what if you don't and you're just making up excuses?"* causing the person to be wide awake in the final panel, their eyes bulging with worry.
- **Restlessness:** This category includes memes that depict physical or mental discomfort, where individuals are unable to sit still or concentrate. Visuals often show characters feeling unsatisfied.
  - Indications: This sub-category is often quite clear, as the meme often contains visuals of someone freaking out, sobbing, or tearing their hair out. The image is usually purposefully blurred or distorted, and the text usually talks about something terrible the person creating the meme had gone through.

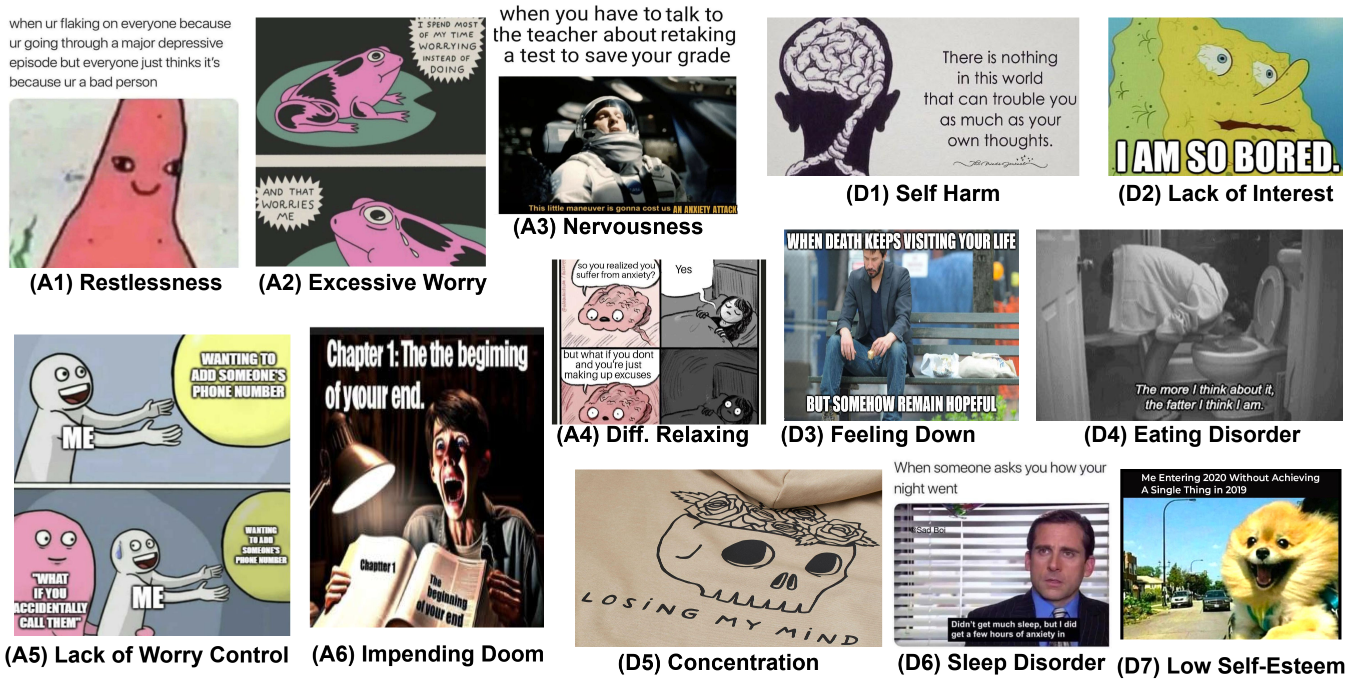


Figure 2: Label-wise sample of memes in the proposed dataset, AxiOM (A1 – A6), illustrating different GAD anxious symptoms and RESTORE (D1 – D7), illustrating different PHQ-9 depressive symptoms.

  - Example: A meme featuring a close-up of Patrick Star with the text *"when ur flaking on everyone because ur going through a major depressive episode but everyone just thinks it's because ur a bad person"* reflects the internal struggle of depression and the external misunderstanding by others.
- **Impending Doom:** This category reflects feelings of looming catastrophe or fear, despite no immediate danger. Memes often depict paranoia or anxiety about something bad happening.
  - Indications: This GAD category may be entailed by a looming presence over someone or someone panicking/worried with text that implies that the person thinks something bad is about to happen in their lives.
  - Example: A meme showing "When you start reading a new book and realize it's predicting your doom" - depicting a terrified person screaming while reading a sinister book titled "The beginning of your end" conveys a sense of impending doom.

RESTORE Dataset. We use an additional public meme depressive symptom classification dataset, RESTORE [45]. The authors carefully compile the dataset using a clinically-informed methodology, drawing from two widely-used social media platforms: *X* and *Reddit*[1]. The original RESTORE categorizes depressive symptoms across eight distinct categories based on the PHQ-9 questionnaire: *Feeling Down, Lack of Interest, Eating Disorder, Sleeping Disorder, Concentration Problem, Lack of Energy, Low Self-Esteem*, and *Self-Harm*. However, upon reviewing the dataset, we observe significant inconsistencies in the annotations associated with the 'Lack of Energy' category. To ensure the integrity of our analysis, we exclude this label from our study, resulting in a focus on the remaining seven categories.

3.2 Data Analysis

AxiOM consists of 3,582 memes categorized into six anxiety-related GAD symptoms, divided into training, validation, and testing sets in a 70:10:20 ratio. Healthcare professionals validated the dataset, making minor adjustments to ensure accurate categorization. Table 1 provides a detailed label-wise analysis, showing the distribution of each anxiety symptom across the splits. The training set, with 2,153 memes, forms the majority, providing a robust foundation for model training. The validation and test sets, containing 307 and 615 memes respectively, ensure balanced fine-tuning and evaluation. This division ensures diverse and consistent representation of anxiety expressions across all GAD categories.

Additionally, we perform a visual analysis to study the relationship between anxiety symptoms and color features, focusing on saturation (vibrancy) and value (brightness). Most memes, regardless of category, are colorful, reflecting heightened and chaotic emotions. Restlessness memes exhibit the highest vibrancy with low hue (red tones), while impending doom and difficulty relaxing memes show the least vibrancy, often featuring grayscale tones, shadows, or evening settings.

Annotator Details, Quality, and Data Ethics. Both datasets, AxiOM and RESTORE, address sensitive mental health topics such as anxiety and depression, and we prioritize ethical considerations at every stage. The annotation process has been conducted by a set of six annotators. Additionally, healthcare professionals verify the quality of each iteration. The annotators represented different age groups,

[1]Reddit: https://www.reddit.com/ ; X (formerly Twitter): https://x.com/

(a) Depression: RESTORE dataset

	LOI	FD	ED	SD	LSE	CP	SH	Total
Train	441	1555	1714	997	549	348	1259	6863
Val	13	55	35	18	29	20	27	197
Test	35	82	61	48	32	50	28	336
Total	489	1692	1810	1063	610	418	1314	7396

(b) Anxiety: AxiOM dataset

	NV	LWC	EW	DR	RST	ID	Total
Train	373	331	322	356	405	366	2153
Val	53	47	46	51	58	52	307
Test	106	94	92	102	116	105	615
Total	532	472	460	509	579	523	3582

Table 1: Data statistics for the depression and anxiety datasets. (a) Depression Labels: LOI: *Lack of interest*; FD: *Feeling down*; ED: *Eating Disorder*; SD: *Sleeping Disorder*; LSE: *Low Self-Esteem*; CP: *Concentration Problem*; SH: *Self Harm* (b) Anxiety Labels: NV: *Nervousness*; LWC: *Lack of Worry Control*; EW: *Excessive Worry*; DR: *Difficulty Relaxing*; RST: *Restlessness*; ID: *Impending Doom*

ranging from 25 to 38, and came from diverse cultural and educational backgrounds, ensuring a broad perspective on mental health symptoms. We source data from publicly available platforms, and all personally identifiable information (PII) is thoroughly anonymized to protect user privacy. Furthermore, we calculate an agreement score of 81% using the Intersection Union Agreement (IUA) method [6], which falls under the strong annotator agreement category.

4 Methodology

In this section, we discuss the complete pipeline of our proposed framework, M3H, and its contributory modules.

Problem Statement. We address a standard multi-modal image-cum-text classification task. Given a meme image (I) from social media consisting of OCR-extracted text (O), the objective is to classify associated symptom (s_i). To achieve this, we fetch and utilize (a) figurative semantics and (b) exemplar knowledge. In the following sections, we detail how to retrieve these insights from memes followed by its integration into M3H for classification.

4.1 Figurative Reasoning

Memes frequently convey layered meanings that are challenging to interpret, even for humans, as they require a deep understanding of commonsense knowledge to figure out the underlying meaning of memes. This knowledge, *aka* figurative knowledge, inherits information related to causality, figurative expression, and cognitive understanding. To address this complexity, we employ prompt engineering on state-of-the-art LLMs to generate commonsense-enriched, figurative reasonings for each meme. We generate these reasonings based on three key attributes that capture different aspects of meme interpretation:

- **Cause-Effect:** Memes often depict scenarios that reflect real-world experiences or situations with clear causes and outcomes. Identifying these cause-effect relationships helps to ground the meme's content in reality, allowing the model to better relate the visual and textual elements to real-life consequences, which are vital for understanding expressions such as anxiety or stress.
- **Figurative Understanding:** Memes are rich in figurative language, often using metaphors, analogies, or symbolic representations to convey deeper, sometimes hidden, messages. Recognizing these figurative elements is essential for decoding the true intent of the meme, which may be cloaked in humor, irony, or satire-common ways users express their struggles indirectly.
- **Mental State:** It is about recognizing the specific psychological state being expressed in the meme.

We employ prompt to fetch this figurative reasoning (r) in order to assist the framework's pipeline in better understanding the meaning and nuances of humor underlying the meme. We present prompt and sample reasoning text in Appendix (c.f Section D).

4.2 Knowledge Fusion

To overcome the limitation of static models being unable to take advantage of dynamic artifacts, we fuse dedicated knowledge into the M3Hthrough a Retrieval-Augmented Generation (RAG) module. This method allows for more contextual and up-to-date meme comprehension and ensures that M3H is informed by the most relevant and recent information. We utilize images (I) in the train split with their figurative reasoning (r) and OCR text (o), $I_t \in \{(o_1, r_1) \dots (o_n, r_n)\}$ to yield most relevant instances from the embedded knowledge, as mentioned below in Equation 1.

$$\mathcal{E} = \Pi([o_1, o_2, \dots, o_n])^{n\times d} \oplus \Pi([r_1, r_2, \dots, r_n])^{n\times d} \tag{1}$$

where Π is a sentence-transformer that yields linear embeddings corresponding to each tuple $\langle o_x, r_x \rangle$. We create the RAG database using the concatenated embeddings of both OCR text and reasoning, resulting in a database of final RAG-Embedding, $\mathcal{E} \in \mathbb{R}^{n\times 2d}$.

For each new meme as an input, we retrieve the top-n most relevant meme indices, γ, from the RAG database. To achieve this, we compute the embedding for the input meme by concatenating the individual embeddings of OCR-text o_k and the figurative content r_k, as shown below.

$$e_k = (\Pi(o_k) \oplus \Pi(r_k)) \in \mathbb{R}^{1\times 2d}; \gamma = \arg\max_i \Phi(e_k, \mathcal{E}_i) \tag{2}$$

The embedding e_k is then compared against the database, $\mathcal{E}$ using cosine similarity Φ, and the output consists of the indices of top-n similar instances, which correspond to the n most similar memes.

4.3 Classifier

Once the relevant examples are retrieved, we combine all the information to construct input for our classifier. The input is designed to guide the classifier in accurately classifying mental health issues depicted in the memes. Our method enhances the understanding of memes through knowledge fusion, and the final classifier leverages text modality to classify symptoms. We present complete prompt

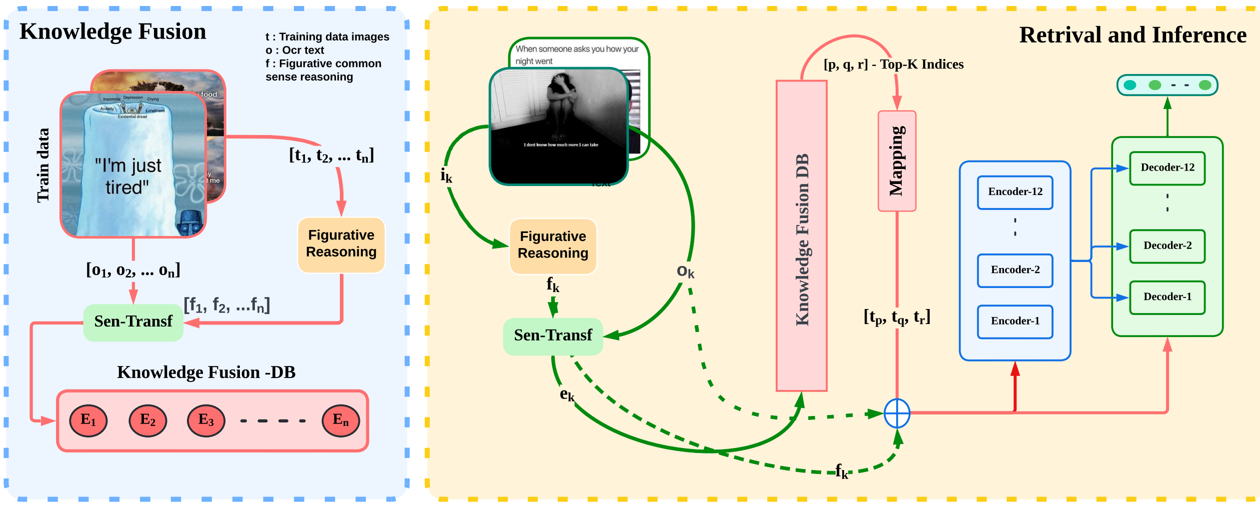


Figure 3: Proposed Framework: M3H. The meme image and OCR text are standalone inputs to our framework. Primarily, we employ LLM for figurative reasoning on prominent commonsense attributes. We deploy the reasoning into a RAG-database to be considered for knowledge fusion during classification. Later, we employ an encoder-decoder framework for final classification.

input with a few examples in the Appendix (c.f. Section D, Table 5). Precisely, we inculcate the following attributes in our final input:

- **OCR-Text** (o): The text extracted from the meme forms an objective component of the classifier's input. This text often includes literal content, such as captions or dialogue within the meme.
- **Figurative Reasoning** (r): Figurative language, as they often use humor, irony, and metaphors to communicate complex states. It generally can be comprehended based on commonsense.
- **Relevant Knowledge Attributes** (γ): It involves retrieving relevant information from the RAG module (knowledge-fusion).

As a result, we employ an encoder-decoder framework that takes the combined input and classifies associated symptoms. This few-shot learning approach, supported by the retrieval of similar examples, enhances the model's ability to generalize from limited data, making it well-suited for analyzing the nuanced nature of memes.

Training. We train the entire framework for two different datasets, namely (a) AxiOM and (b) RESTORE. Here, the primary dataset possesses a multiclass symptom classification problem, whereas the latter is a multilabel setting. For M3H to accommodate this requirement, we put a classifier personalized to each of these problem statements. For AxiOM, we put a single softmax for a single class. On the other hand, for RESTORE, the classification for each class y_j is determined as follows:

$$y_j = \begin{cases} 1 & \text{if } P(c_j \mid o_k, r_k, \gamma_k) > \tau_j, \\ 0 & \text{if } P(c_j \mid o_k, r_k, \gamma_k) \leq \tau_j, \end{cases} \tag{3}$$

where γ_k represent the set of top-n similar examples, and we set τ_j as threshold for multilabels.

5 Experiments, Results, and Analysis

Here, we discuss the selection of baselines and compare their performance with the proposed method, M3H. Further, we present an exhaustive human evaluation and analysis.

5.1 Baselines

In our work, we experiment with the following LLMs for figurative reasoning: ►**COMET** [4]: Commonsense transformer with better contextual understanding of textual information by capturing commonsense relationships. ►**LLAVA** [22]: It is a multimodal LLM pretrained on textual and visual data providing better cross-modal modeling. In our case, it is observed to understand complex figurative nuances well. ►**GPT-4o** [30]: GPT-4o is state-of-the-art LLM capable of handling both textual and visual data. It performs best due to its ability to interpret multimodal and diverse nuances. Next, we experiment with the following classification heads within our framework: ►**BERT** [8]: A widely known encoder in NLP literature, often compared with other models like DeBERTa. ►**DeBERTa** [13]: A notable extension of BERT, incorporating disentangled attention and more advanced techniques for improved contextual understanding. ►**Mental-BERT** [15]: A BERT-based model specifically trained on mental health datasets adapted for tasks in this domain. ►**BART** [21]: A powerful encoder-decoder model used for generative tasks in NLP. ►**Mental-BART** [46]: A variation of BART, fine-tuned on mental health corpora to handle domain-specific challenges. ►**Yadav et al. [45]** ($SOTA$): This baseline implementation focuses on generating text-image feature representations, aligning visual and textual tokens, currently the prior state-of-the-art.

Evaluation Metric. We present our comparative evaluation using F1 scores (Macro-F1, Weighted-F1, and Micro-F1) on seven classes,

Models	Depressive Memes		Anxiety Memes	
	Macro-F1	Weigh-F1	Macro-F1	Weigh-F1
OCR-Text:				
BERT	58.16	57.12	62.99	63.06
Deberta-V3	57.99	57.40	62.03	61.86
Mental-BERT	59.14	60.03	63.41	63.40
BART	41.56	38.35	62.82	62.62
Mental-BART	54.19	52.55	64.97	64.88
OCR + COMET (f_k):				
BERT	58.71	58.14	61.00	61.00
Deberta-V3	58.81	58.90	60.52	60.17
Mental-BERT	60.11	60.79	63.00	63.00
BART	62.55	63.28	62.71	62.61
Mental-BART	62.94	63.43	63.97	63.86
OCR + LLAVA (f_k):				
BERT	63.97	66.15	62.43	62.38
Deberta-V3	60.80	60.55	60.70	60.80
Mental-BERT	65.45	63.02	63.00	63.00
BART	67.00	68.55	64.01	64.02
Mental-BART	65.06	66.00	65.03	65.07
OCR + GPT-4o (f_k):				
BERT	58.65	59.05	62.01	62.02
Deberta-V3	59.87	59.61	61.33	62.40
Mental-BERT	60.80	60.29	62.38	62.52
BART	63.32	64.09	64.00	64.05
Mental-BART	63.26	64.82	65.51	65.54
Yadav et al. [45] (*SOTA*):	63.58	64.59	64.21	65.34
M3H (*Proposed*):	**67.52**	**68.79**	**70.00**	**70.10**
– (FCS+RAG)	65.20	66.75	64.28	64.97
– (OCR+RAG)	66.06	66.01	62.54	62.53
– FCS	58.50	58.91	62.82	62.62
$\Delta_{M3H-SOTA}(\%)$	↑ 4.94	↑ 4.20	↑ 5.79	↑ 4.66

Table 2: `M3H`'s Performance and Ablation Study. `M3H` clearly outperforms all 20 variations of baseline across both datasets: `AxiOM` and `RESTORE` on both macro- and weighted-F1. It is worth noting that we perform multilabel classification on RESTORE, unlike on `AxiOM`, which is standard multiclass classification. Due to space constraints, we enclose additional ablations and metrics in Appendix (c.f. Table 4).

demonstrating that our proposed approach outperforms these baselines across the board. To make the results consistent, we ran Yadav et al. [45] model on `AxiOM` and tabulated scores in the Table 2.

5.2 Performance Comparison

Our proposed model, `M3H`, shows substantial improvements across 20 variations of baselines, as shown in Table 2. Primarily, we experiment with standalone OCR text, where Mental-BERT performed best with a score of 59.14% and 60.03% on `AxiOM`, whereas on `RESTORE`, it yields 63.41% and 63.40% on macro- and weighted-F1, respectively. Next, we compare the performance of commonsense reasoning in order to extract figurative nuances. We compare COMET, LLaVA, and GPT-4o and observe that GPT-4o performed better across both datasets. On `AxiOM`, we observe an increment of +7.06% and +12.90% with BART and Mental-BART; whereas, on RESTORE, we observe a marginal improvement of +0.83% and +1.01% with Mental-BART on macro- and weighted-F1 metrics, respectively. As a result, we fuse `M3H` to consider figurative commonsense nuances from `GPT-4o` and classifier as BART, which

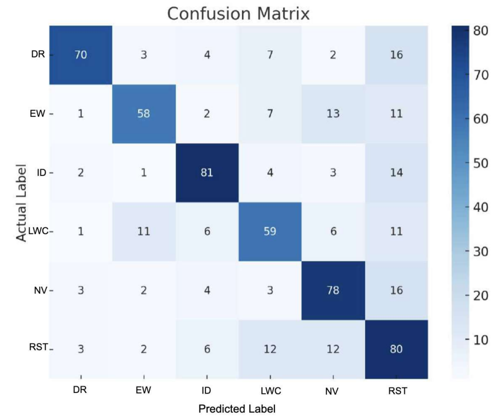


Figure 4: Confusion matrix computed on our proposed dataset, `AxiOM` using our model, `M3H`.

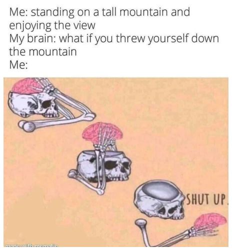
provides the best output. With this setting, `M3H` achieves F1 scores of 67.52 (+494%) and 68.79 (+4.20%), and 70.09 on macro-, weighted- and micro-F1. On the other hand, we observe a consistent improvement on RESTORE data as well, with 70.00% score on both macro and weighted-F1 score, with an improvement of +5.79% and +4.66%, respectively. These results highlight `M3H`'s superiority in classifying symptoms. We present additional metrics and results in the Appendix (c.f. Table 4).

Ablation Study. To better understand the contributions of each component in `M3H`, we conduct an ablation study. We assess the impact of each module on the model's performance, as shown in Table 2. Our findings indicate a significant decline in `M3H`'s performance upon the removal of the figurative reasoning and knowledge module. Specifically, the weighted-F1 score for the depression dataset dropped from 68.79 to 65.20 when the FCS-RAG was removed. Additionally, the score further decreased to 66.06 with the removal of the OCR-RAG. A similar trend is observed in the `AxiOM`, where macro-F1 score falls from 70.10 to 64.97 following the removal of the FCS-RAG and further falls to 62.53 upon the removal of the OCR-RAG. Notably, removing the FCS module results in scores of 58.91 and 62.62 for the `RESTORE` and `AxiOM` datasets, respectively, showing that both modules contribute positively. We further observe an increment of +5.09% to quantify the contribution of the RAG module (c.f. Appendix; Figure 7). We present additional ablations in the Appendix due to space constraints (c.f. Section C and Table 4).

5.3 Error Analysis

In this section, we present a two-way error analysis of `M3H` performance through both quantitative and qualitative aspects.

Quantitative Analysis. We report the confusion matrix for our model in 4. We observe several pairs with misclassification ($\geq$ 15%). The most significant confusion is between the classes 'Lack of Worry Control' and 'Impending Doom', where 16% of the former were misclassified as 'Impending Doom'. This confusion can be attributed to the similarity in the emotional expression of the two classes, particularly in situations where anxiety involves intrusive thoughts, which is hard to understand casually for even humans. We observe a similar confusion between 'Restlessness' vs 'Lack of



	Predictions	True
Figurative Reasoning: - **Cause-Effect:** Standing on a tall mountain and enjoying view. Triggering intrusive thoughts about self-harm - **Figurative Understanding:** The mountain represents a moment of peace or happiness. The exaggerated response of the brain contrasts with the calmness of enjoying a view, emphasizing the irony of anxiety. - **Mental State:** Frustration and dismissal of harmful thoughts.	Impending Doom	Lack of Worry Control
Figurative Reasoning: - **Cause-Effect:** Therapist suggests being more social. Individual humorously invites people over for a "crying session." - **Figurative Understanding:** "Crying session" symbolizes emotional vulnerability. - **Mental State:** Individual feels overwhelmed or anxious about social interactions.The invitation to a "crying session" suggests feelings of sadness	Impending Doom	Lack of Worry Control

Figure 5: Error Analysis. We attempt to connect the dots between misclassification and figurative reasoning. Evidently, the reasoning module is able to capture underlying nuances; however, we observe label confusion due to which misclassification happens.

Worry Control', with a confusion rate of 12.2%. Evidently, M3H tends to confuse the emotional content, where the individual is struggling to manage their emotions, suggesting other labels.

Qualitative Analysis. We examine the performance of M3H and observe that the model sometimes misinterprets the humor or figurative undertones of the text. We present a few such cases in Figure 5. For instance, in case #1, an individual standing on a tall mountain expresses an intrusive thought about jumping. The correct label is 'Lack of Worry Control', reflecting the person's struggle to manage these intrusive thoughts; however, the model predicted 'Impending Doom' as a symptom likely due to the high-stakes nature of the situation. In case #2, a meme about a 'crying session' is labeled as 'Lack of Worry Control', but the model predicts 'Restlessness'. The underlying sarcasm, where the individual uses a 'crying session' to cope with social anxiety, indicates emotional difficulty rather than physical restlessness; however, M3H focuses on the anxious behavior implied by the invitation, resulting in misclassification.

5.4 Human Evaluation

To assess the quality of the commonsense reasoning generated by GPT-4 for both anxious and depressive memes, we conduct an human evaluation for both datasets. We select a total of 40 random samples, 20 from each dataset. Each meme sample includes figurative reasoning that includes three attributes: *cause and effect*, *figurative understanding*, and *mental state*. We ask human evaluators to rate the commonsense reasoning on the likert scale of 1 (poor) to 5 (excellent). Evidently, as shown in Figure 6, the majority of feedback in both datasets falls between 4 and 5, indicating that the figurative-cum-commonsense reasoning is generally well-perceived, aligning with human interpretations of the underlying sarcasm and mental health context.

6 Discussion, Ethics, and Limitations

Our findings emphasize the importance of integrating visual and textual commonsense reasoning to accurately classify depressive and anxious symptoms in memes. While M3H demonstrates strong performance, there is room for improvement. Relying solely on online memes limits the framework's ability to capture complex, trendy nuances due to a lack of diverse knowledge. Although LLMs enhance performance through commonsense reasoning, they increase computational complexity and may produce responses misaligned with human reasoning or fail to capture meme-specific nuances, which are critical in mental health contexts. Importantly, our framework is not intended to diagnose mental health disorders but to analyze social media posting behavior related to depressive and anxious meme content. Future work will focus on improving the diversity, factuality, and faithfulness of generated reasoning. Expanding our approach to include other domains of figurative communication, such as political cartoons or digital art, could provide a broader understanding of memes and their contextual issues.

7 Conclusion

Social media has become a prominent platform for expressing mental health issues through memes. In this work, we present M3H, a novel framework that leverages GPT-4 for figurative reasoning and a BART-based transformer for identifying symptoms depicted in memes. Through rigorous benchmarking against six competitive baselines (20 variations), including state-of-the-art models, we demonstrate M3H's superiority on both the newly proposed AxiOM dataset for anxious symptom classification and the RESTORE dataset for depressive symptom classification. Using macro-F1 and weighted-F1 scores, M3Hachieves significant improvements of 4.94% and 5.79% in macro-F1, and 4.20% and 4.66% in weighted-F1 across both datasets. Our study is supported by exhaustive human evaluation, qualitative and quantitative analyses, and a discussion on ethics and limitations.

Acknowledgment

The authors acknowledge the support of the Infosys Foundation through CAI at IIIT-Delhi and the iHub-Anubhuti-iiitd Foundation.

References

[1] Umair Akram and Jennifer Drabble. 2022. Mental health memes: beneficial or aversive in relation to psychiatric symptoms? *Humanities and Social Sciences Communications* 9, 1 (2022), 1–6.

[2] Ankita Bhaumik and Tomek Strzalkowski. 2024. Towards a Generative Approach for Emotion Detection and Reasoning. *arXiv preprint arXiv:2408.04906* (2024).

[3] Sravani Boinepelli. 2022. *Towards Identification, Classification and Analysis of Mental Illness on Social Media*. Ph. D. Dissertation. International Institute of Information Technology, Hyderabad.

[4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. arXiv:1906.05317 [cs.CL] https://arxiv.org/abs/1906.05317

[5] Luca Braghieri, Ro'ee Levy, and Alexey Makarin. 2022. Social media and mental health. *American Economic Review* 112, 11 (2022), 3660–3693.

[6] Pankaj K Choudhary and HN Nagaraja. 2005. Assessment of agreement using intersection-union principle. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 47, 5 (2005), 674–681.

[7] Munmun De Choudhury. 2013. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*. 49–52.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[9] Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 153–164.

[10] Myryll Esco, Kenneth Llanasas, Vida Zoe Salino, Carmel Guazon, and Lucy L Teves. 2023. Perceived Attitude of Facebook Users Towards Memes About Depression. (2023).

[11] Aastha Gupta, Deepanshi Gupta, Pallavi Chaudhary, Pragya Raje, and Komal Chandiramani. 2021. Exploring trivialisation of mental health issues using internet memes in young adults. *The* (2021), 868–81.

[12] Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Amanullah Yasin, Gautam Srivastava, Dawid Połap, Thippa Reddy Gadekallu, and Zunera Jalil. 2024. Multimodal religiously hateful social media memes classification based on textual and image data. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 8 (2024), 1–19.

[13] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654 [cs.CL] https://arxiv.org/abs/2006.03654

[14] Blake Anthony Hickey, Taryn Chalmers, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. 2021. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors* 21, 10 (2021), 3461.

[15] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. arXiv:2110.15621 [cs.CL] https://arxiv.org/abs/2110.15621

[16] Kimesha L Jones. 2022. *Perceptions of Facebook by Individuals with Mental Health Vulnerabilities*. Northcentral University.

[17] Saurav Joshi, Filip Ilievski, and Luca Luceri. 2024. Contextualizing Internet Memes Across Social Media Platforms. In *Companion Proceedings of the ACM on Web Conference 2024*. 1831–1840.

[18] Abdul Aziz Turhan Kariko and Nonny Anasih. 2019. Laughing at one's self: A study of self-reflective internet memes. In *Journal of physics: conference series*, Vol. 1175. IOP Publishing, 012250.

[19] Naseer Ahmed Khan, Samer Abdulateef Waheeb, Atif Riaz, and Xuequn Shang. 2021. A novel knowledge distillation-based feature selection for the classification of ADHD. *Biomolecules* 11, 8 (2021), 1093.

[20] Alexander V Laskin and Yasha A Laskin. 2022. Memes: Interactive Creative Intertextuality on Social Media. *The SAGE Handbook of Social Media Marketing* (2022), 307.

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL] https://arxiv.org/abs/1910.13461

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV] https://arxiv.org/abs/2304.08485

[23] Usha Lokala, Aseem Srivastava, Triyasha Ghosh Dastidar, Tanmoy Chakraborty, Md Shad Akhtar, Maryam Panahiazar, and Amit Sheth. 2022. A Computational Approach to Understand Mental Health from Reddit: Knowledge-Aware Multitask Learning Framework. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 640–650. https://doi.org/10.1609/icwsm.v16i1.19322

[24] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and Time-aware Joint Contextual Learning for Dialogue-act Classification in Counselling Conversations *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 735–745. https://doi.org/10.1145/3488560.3498509

[25] Suresh Malodia, Amandeep Dhir, Anil Bilgihan, Pranao Sinha, and Tanishka Tikoo. 2022. Meme marketing: How can marketers drive better engagement using viral memes? *Psychology & Marketing* 39, 9 (2022), 1775–1801.

[26] Laura Manea, Simon Gilbody, and Dean McMillan. 2015. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *General hospital psychiatry* 37, 1 (2015), 67–75.

[27] Elly Atiqah Md Nawawi, Muhammad Faiz Izzudin, Nur Atiqah Mohd Zulkafli, and Siti Sarah Januri. 2020. Identifying the Factors Affecting Internet Memes to Become Viral on Social Media. (2020).

[28] Saif M Mohammad. 2024. WorryWords: Norms of Anxiety Association for over 44k English Words. *arXiv preprint arXiv:2411.03966* (2024).

[29] Jessica Gall Myrick, Robin L Nabi, and Nicholas J Eng. 2022. Consuming memes during the COVID pandemic: Effects of memes and meme type on COVID-related stress and coping efficacy. *Psychology of Popular Media* 11, 3 (2022), 316.

[30] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[31] Umera Wajeed Pasha. 2024. Multilingual sexism detection in memes, a CLIP-enhanced machine learning approach. *Working Notes of CLEF* (2024).

[32] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 4439–4455. https://doi.org/10.18653/v1/2021.findings-emnlp.379

[33] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184* (2021).

[34] Adam G Sanford, Stacy L Smith, and Dinur Blum. 2023. Going viral: how social media increased the spread of COVID-19 misinformation. In *COVID-19: Individual Rights and Community Responsibilities*. Routledge, 54–76.

[35] Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274* (2022).

[36] E. S. Smitha, S. Sendhilkumar, and G. S. Mahalaksmi. 2018. Meme Classification Using Textual and Visual Features. In *Computational Vision and Bio Inspired Computing*, D. Jude Hemanth and S. Smys (Eds.). Springer International Publishing, Cham, 1015–1031.

[37] Aseem Srivastava, Tanya Gupta, Alison Cerezo, Sarah Peregrine (Grin) Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2025. Critical behavioral traits foster peer engagement in Online Mental Health Communities. *PLOS ONE* 20, 1 (01 2025), 1–17. https://doi.org/10.1371/journal.pone.0316906

[38] Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, and Md Shad Akhtar. 2024. Knowledge Planning in Large Language Models for Domain-Aligned Counseling Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17775–17789. https://doi.org/10.18653/v1/2024.emnlp-main.984

[39] Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the Shared Task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Parameswari Krishnamurthy, and Elizabeth Sherly (Eds.). Association for Computational Linguistics, Kyiv, 126–132. https://aclanthology.org/2021.dravidianlangtech-1.16

[40] Viriya Taecharungroj and Pitchganut Nueangjamnong. 2014. The effect of humour on virality: The study of Internet memes on social media. In *7th International Forum on Public Relations and Advertising Media Impacts on Culture and Social Communication. Bangkok, August*. 151–166.

[41] Sorina Vasile, Lebogang Mototo, and Tinashe Chuchu. 2021. Using "Memes" as a marketing communication tool in connecting with consumers within the age of digital connectivity. *International Review of Management and Marketing* 11, 3 (2021), 30–35.

[42] Fan Wu, Guolian Chen, Junkuo Cao, Yuhan Yan, and Zhongneng Li. 2024. Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism. *Electronics* 13, 14 (2024). https://doi.org/10.3390/electronics13142780

[43] Lexing Xie, Apostol Natsev, John R Kender, Matthew Hill, and John R Smith. 2011. Visual memes in social media: tracking real-world news in youtube videos. In *Proceedings of the 19th ACM international conference on Multimedia*. 53–62.

[44] Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. 2023. Towards identifying fine-grained depression symptoms

from memes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8890–8905.
[45] Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. 2023. Towards Identifying Fine-Grained Depression Symptoms from Memes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 8890–8905. https://doi.org/10.18653/v1/2023.acl-long.495
[46] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023. MentalLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. arXiv:2309.13567 [cs.CL]
[47] Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal Understanding of Memes with Fair Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2007–2017.
[48] Dawei Zhou, Jiebo Luo, Vincent Silenzio, Yun Zhou, Jile Hu, Glenn Currier, and Henry Kautz. 2015. Tackling mental health by integrating unobtrusive multimodal sensing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29.

A Data Collection

We gather data from various online sources, including Reddit and Instagram, which are popular platforms known for unfiltered expression and meme sharing. The anonymity provided by these platforms allows for more expressive meme content. To collect data from Reddit, we use Reddit's API and developed a scraper to extract images from specific subreddits, setting a maximum limit of 3,000 images per subreddit. For Instagram, we applied a similar approach to extract images from public profiles using our scraper. The subreddits, such as *r/2meirl4meirl, r/anxietymemes, r/healthanxietyhumor, r/sa_memetherapy, r/socialanxietymemes*, and *r/TrollCoping*, were selected based on Google search results identifying popular mental health subreddits with frequent meme-posting about mental health struggles. These subreddits capture diverse anxiety-related expressions, including depression-related content (e.g., impending doom and restlessness from sleep disturbances in *r/2meirl4meirl*) and raw, unfiltered memes reflecting anxious characteristics like dissociation and distorted reality (e.g., *r/bpd_memes*). For Instagram, we selected profiles like *anxietian, anxietytwitter*, and *anxietymemetherapy*, using the WorryWords lexicon [28], a comprehensive list of over 44,000 English words scored for their association with anxiety. Profiles were chosen based on the presence of anxiety-related keywords in their usernames, bios, or handles. The collected memes were annotated with their expressed mental health symptoms based on a predefined set of symptoms. This approach ensures a broad spectrum of anxiety-related expressions, enriched by contributions from individuals with varied backgrounds and mental health conditions.

B Experimental Setup

For the baseline's training, we have inferred all the models from Hugging Face and trained for 10 epochs. It took approximately 50 minutes for BERT-based models and 1.2 hours for BART-based models, using 3 NVIDIA Tesla V100 GPUs. For the proposed model, we have trained for 10 epochs, with each epoch taking about 30 minutes. The inference time for classifying a depression category is approximately 5 seconds. All training and inference operations were conducted on 3 NVIDIA Tesla V100 GPUs.

We employed the AdamW optimizer with a learning rate of 510¯5, beta1 of 0.9, and beta2 of 0.999. To prevent overfitting, we applied a dropout rate of 0.2 and a weight decay of 110¯2. For single-label classification tasks, we used Cross Entropy Loss, while Binary Cross Entropy Loss was employed for multi-label scenarios. The model's performance was evaluated using Micro, Macro, and Weighted F1 Scores to provide a comprehensive assessment across different class distributions and task types.

The learning rate scheduler was set to constant, and we used a batch size of 16. The RMS Norm Epsilon was set to 110¯5, and the Adam Epsilon to 110¯8. The maximum sequence length was capped at 512 tokens. Gradient clipping was applied with a threshold of 1.0 to stabilize training.

These hyperparameters were carefully tuned to optimize model performance while balancing computational efficiency. The detailed hyperparameters are provided in Table 1, which includes additional parameters such as warmup steps and specific optimizer settings.

This configuration allowed us to effectively train and evaluate our models across various mental health meme classification tasks, accounting for both single-label and multi-label scenarios, while ensuring robust performance metrics through the use of diverse F1 score calculations.

Hyperparameter	Value
Learning Rate (lr)	5×10^{-5}
Adam Beta1	0.9
Adam Beta2	0.999
Adam Epsilon	1×10^{-8}
RMS Norm Epsilon	1×10^{-5}
Dropout	0.2
Batch Size	16
Learning rate schedular	constant
Lora Rank (r)	16
Lora alpha (α)	8
Target modules	W_q, W_k, W_v, W_o

Table 3: Hyperparameters used in the experiment.

C Additional Ablation Study

Each component of the M3H module plays a crucial role in enhancing performance, as demonstrated in Table 4. The table outlines how removing one or more components affects the overall model performance. Here, OCR refers to Optical Character Recognition of text from the meme, while **FCS** (Figurative Common Sense) represents deep understanding extracted from multimodal large language models. We report **Macro F1**, **Weighted F1**, and **Micro F1** scores on both the **RESTORE** and **AxioM** datasets. From the ablation results, we observe that incorporating all components yields the highest scores across all metrics. On the other hand, in Figure 7, the performance improvement with each additional component is visualized. Notably, the system exhibits a significant performance boost of ΔRAG = 5.40 after introducing knowledge-fusion, indicating its effectiveness in enhancing the system's overall capability.

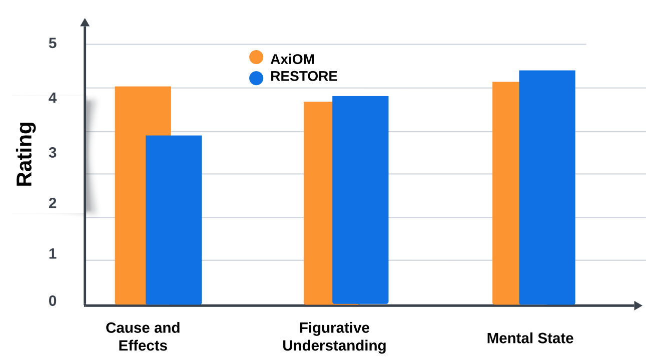


Figure 6: Human Evaluation for commonsense reasoning using GPT-4o for each figurative reasoning attribute on both dataset: AxiOM and RESTORE.

OCR	OCR-RAG	FCS	FCS-RAG	Depression			Anxiety		
				Macro-F1	Weighted-F1	Micro-F1	Macro-F1	Weighted-F1	Micro-F1
✓	✗	✗	✗	58.50	58.91	62.29	62.82	62.62	62.21
✓	✓	✗	✗	66.01	67.33	68.60	62.54	62.53	62.05
✗	✗	✓	✗	61.18	62.05	66.84	57.00	58.01	56.84
✗	✗	✓	✓	66.42	67.65	69.24	60.45	60.69	60.46
✓	✗	✓	✗	63.32	64.09	66.22	64.28	64.97	64.20
✓	✗	✓	✓	65.20	66.75	69.01	64.94	64.95	64.84
✓	✓	✓	✗	66.06	66.01	67.92	61.82	61.79	62.21
✓	✓	✓	✓	**67.52**	**68.79**	**70.09**	**70.00**	**70.00**	**69.38**

Table 4: Experimental results for various model configurations

D Prompts and Examples

Analyze the following anxiety meme image to extract common sense reasoning in the form of triples.. These relationships should capture the following elements:

- 1. Cause-Effect: Identify concrete causes or results of the situation depicted in the meme.
- 2. Figurative Understanding: Capture underlying metaphors, analogies, or symbolic meanings that convey the meme's deeper message, including any ironic or humorous undertones.
- 3. Mental State: Capture specific mental or emotional states depicted in the meme.

Analyze the following depression meme image to extract common sense reasoning in the form of triples. These relationships should capture the following elements:

- 1. Cause-Effect: Identify concrete causes or results of the situation depicted in the meme.
- 2. Figurative Understanding: Capture underlying metaphors, analogies, or symbolic meanings that convey the meme's deeper message, including any ironic or humorous undertones.
- 3. Mental State: Capture specific mental or emotional states depicted in the meme.

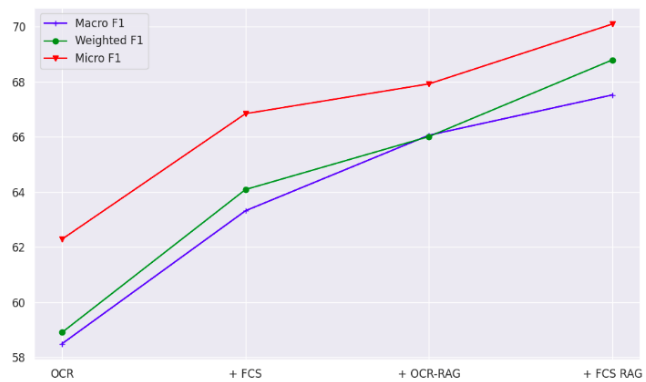


Figure 7: There is a significant increase in the score when RAG is added into the system Δ RAG = 5.40 % on Depressive meme dataset.

System prompt: You specialize in analyzing mental health behaviors through social media posts. Your task is to classify the mental health issue depicted in a person's post from the following categories: ['Difficulty Relaxing', 'Excessive Worry', 'Impending Doom', 'Lack of Worry Control', 'Nervousness', 'Restlessness'].

Example-1:
<|ocr_text|> "I DONT ALWAYS SLEEP AT NIGHT BUT WHEN I DO, I DONT"
<|commonsense figurative explanation|> Here are some common sense reasoning triples extracted from the anxiety meme:
Cause-Effect
1. Cause: Not sleeping regularly at night
   Effect: Feeling tired or anxious during the day.
2. Cause: Drinking alcohol (implied with the beer bottle)
   Effect: Difficulty in sleeping or restlessness at night.

Figurative Understanding
1. Metaphor: "I don't always sleep at night"
   Interpretation: Suggests a chaotic or irregular sleep pattern that many can relate to, often due to anxiety.
2. Irony: "But when I do, I don't"
   Interpretation: Implies that even when he does sleep, it's not restful, highlighting the contradiction in seeking rest yet feeling unrested.

Mental State
1. Mental State: Anxiety or restlessness at night
   Description: Reflects the inner turmoil of being unable to sleep due to worries or stress.
2. Mental State: Humor or sarcasm
   Description: The playful tone of the statement conveys a sense of humor about the struggles of insomnia or disrupted sleep patterns.

<|Assistant|> The mental health disorder of the person for this post is: **Difficulty Relaxing**

Example-2:
<|ocr_text|> "am I exhausted? Yes will I go to bed earlier? No"
<|commonsense figurative explanation|> Here are some triples that capture the common sense reasoning from the anxiety meme image:
Cause-Effect:
1. Cause: The individual feels exhausted.
2. Effect: Despite this exhaustion, the individual will not go to bed earlier.

Figurative Understanding:
1. Symbolic Meaning: The exhaustion represents a common struggle with prioritizing self-care.
2. Humorous Undertone: The light-hearted refusal to go to bed earlier despite acknowledgment of exhaustion illustrates a relatable, ironic attitude towards personal habits.

Mental State:
1. Mental State: The individual is feeling exhausted.
2. Emotional State: There is a sense of defiance or acceptance of unhealthy habits, possibly mixed with humor or resignation.

<|Assistant|> The mental health disorder of the person for this post is: **Difficulty Relaxing**

Your turn:
<|ocr_text|> "Dad: you're up early" - "Me:" - "Dad: you never went to sleep did you"
<|commonsense figurative explanation|> Here are the extracted triples based on the anxiety meme:
Cause-Effect
1. Cause: The child is up early.
   Effect: The dad questions the child's sleep pattern.
2. Cause The child did not go to sleep.
   Effect: The dad realizes the child may be sleep-deprived or acting unusually.

Figurative Understanding:
1. Figurative Meaning The dad's smile symbolizes an awareness of common teenage behavior (e.g., staying up late).
2. Underlying Message: It humorously highlights a typical parent-child dynamic regarding sleep schedules.
3. Irony The child being up early might typically be seen as positive, but it signifies a lack of sleep instead, turning a seemingly good situation into a concerning one.

Mental State:
1. Mental State of the Dad: Amusement or bemusement at the child's situation.
   Emotion: Curiosity mixed with a hint of concern.
2. Mental State of the Child Possibly anxious or guilty about not sleeping.
   Emotion: Discomfort or embarrassment due to the revelation of staying awake all night.

<|Assistant|> The mental health disorder of the person for this post is: || M3H Prediction: **Difficulty Relaxing**

Table 5: This table presents the detailed end-to-end prompt used for our proposed model, M3H, along with the corresponding prediction.