# Mental Health Meme Classification

**Aman Chauhan**
MT23015

**Megha**
MT23125

## 1 Introduction

The rise of internet memes as a cultural medium has transformed how individuals, particularly younger audiences, express and engage with mental health experiences. Platforms like Reddit and Instagram host a wide variety of memes that subtly communicate emotional distress, often through sarcasm, metaphors, irony, or symbolic imagery. For instance, a meme with the caption "I don't know how much more I can take" overlaid on a distressed cartoon figure conveys emotional pain in a non-literal, figurative manner.

While humans can effortlessly interpret such content using commonsense knowledge and figurative reasoning, most current Multimodal Language Models (MLMs) struggle to grasp the nuanced emotional states embedded within these memes. This gap arises because MLMs often lack the ability to understand the cause-effect relationships, the underlying mental state, and the figurative expressions encoded in both the visual and textual elements of memes.

To address this challenge, the Fuse-MH framework (Figurative-Understanding and Semantic Embedding for Mental Health memes) was proposed. It combines Visual-Language Models (VLMs) for processing image-text pairs, retrieval-augmented generation (RAG) to incorporate external commonsense and affective knowledge, and large language models like GPT-4o for figurative interpretation. This enriched understanding is passed to a BART-based classifier that categorizes the meme into fine-grained mental health symptom classes.

The paper introduces the AxiOM dataset, comprising 3,582 anxiety-related memes, annotated with six clinically grounded anxiety symptoms-such as restlessness, excessive worry, and lack of worry control-derived from the Generalized Anxiety Disorder (GAD) questionnaire. Expert-verified annotations ensure high quality and alignment with real-world mental health assessment standards.

This project builds upon the motivation that meme-based communication of mental health is a growing trend. Yet, the interpretive limitations of AI pose significant barriers to supporting users experiencing distress. Our work focuses on understanding and classifying such memes with a hybrid architecture that bridges visual and textual semantics, draws on external knowledge, and recognizes the role of figurative language in online mental health discourse.

## 2 Related Work

- **Text-Based Mental Health Detection:** Traditional research in automated mental health detection has largely focused on textual data from platforms like Twitter, Reddit, and forums. Techniques based on Natural Language Processing (NLP) and Machine Learning (ML) have been used to identify linguistic patterns associated with depression, anxiety, and suicidal ideation. Early works utilized features such as word frequency, sentiment, and psycholinguistic cues (e.g., LIWC), while more recent approaches have adopted deep learning models, such as LSTMs, CNNs, and transformer-based architectures like BERT and RoBERTa, to capture contextual semantics and perform binary or multi-class classification of mental health symptoms.

- **Multimodal Approaches and the Rise of Memes:** With the evolution of visual-first platforms like Instagram and Tumblr, researchers have begun exploring multimodal mental health detection, especially through memes—rich in both images and figurative captions. Memes often convey complex emotional content using sarcasm, symbolism, and humor, which are challenging for traditional AI models to interpret. Some prior studies

have explored extracting visual features using CNN-based encoders (e.g., ResNet) and combining them with textual embeddings, but many of these approaches still underperform due to the figurative and commonsense-heavy nature of memes.

- **Commonsense and Figurative Reasoning in Mental Health Meme Classification:** Mazhar et al. (2025) proposed a novel approach for classifying mental health memes by leveraging both figurative language and commonsense knowledge. They utilized transformer-based models, specifically COMET, to capture and integrate implicit meanings, such as metaphors and idioms, from the text. This approach helped the model better understand the figurative language commonly used in mental health-related memes. In their work, they focused on text-based classification, and did not explicitly incorporate image features for classification, relying instead on the powerful capabilities of transformer models to process and interpret complex text patterns. This combination of commonsense knowledge and figurative reasoning enhanced the model's ability to accurately classify memes related to depression and anxiety.

**Gaps and Motivations:** While Mazhar et al.'s work highlights the importance of figurative understanding and commonsense reasoning, it leaves open opportunities in directly incorporating image features, especially in memes where visual context is crucial to interpreting emotional or humorous intent. Our work aims to build on this foundation by leveraging image-text fusion, sarcasm-target integration, and shared feature fusion mechanisms to better model the multimodal nature of mental health memes.

## 3 Dataset

- **Depression Dataset:** contains 9837 meme images with corresponding labels in json file of 8 classes:
  - **Eating Disorder** - Memes reflecting unhealthy eating behaviors such as overeating or appetite loss, often as coping mechanisms.
    *Example:* A character sitting on the floor surrounded by junk food wrappers: I'll just eat a snack… 12 hours later: empties fridge.
  - **Self-Harm** - Content hinting at physical self-injury or ideation, depicted metaphorically or explicitly.
    *Example:* A black-and-white image of a person with a shadow shaped like a blade. Text: "Pain is the only thing that feels real"
  - **Feeling Down** - Expressions of persistent sadness, low mood, or hopelessness.
    *Example:* A happy-face mask held in front of a sad cartoon face. Text: "Smiling outside, dying inside."
  - **Low Self-Esteem** - Memes showcasing negative self-image or feelings of worthlessness.
    *Example:* A figure in the mirror showing a broken reflection. Text: "Why would anyone like someone like me?"
  - **Lack of Energy** - Indications of physical or emotional fatigue, difficulty in initiating tasks.
    *Example:* A person in pajamas asleep at their work desk with zzz bubbles. Text: "Me waking up tired from my 14-hour nap"
  - **Sleeping Disorder** - Themes of insomnia, oversleeping, or disrupted sleep patterns.
    *Example:* A dark room with someone wide awake in bed, eyes wide open. Text: "Brain: Let's think about that one thing you did 5 years ago at 3am."
  - **Concentration Problem** - Difficulty focusing, mental fog, or distractibility often portrayed humorously.
    *Example:* An open book with doodles floating out of it; character looks dazed. Text: "Me reading one sentence 10 times and still not getting it"
  - **Lack of Interest** - Reduced engagement or loss of pleasure in usual activities, commonly conveyed through passive or detached visuals.
    *Example:* A grayscale meme with a slouched figure staring at a phone. Text: "Everything's boring. Even memes."
- **Anxiety Dataset:** contains 3260 meme images with corresponding labels in json file of

6 classes:

- **Restlessness** - Memes suggesting inability to relax, fidgeting, or being on edge.
  *Example:* A figure pacing in circles, wearing out the floor. Text: "Why sit still when you can pace around for 2 hours?"
- **Lack of Worry Control** - Depictions of overwhelming worry that feels uncontrollable or intrusive.
  *Example:* A brain surrounded by tangled wires and alerts. Text: "Worrying about worrying too much"
- **Nervousness** - Indicators of heightened tension, fear, or unease in everyday scenarios.
  *Example:* A sweating figure looking at phone with shaky hands. Text: "When you say 'hi' and they don't respond immediately"
- **Impending Doom** - Expressions implying catastrophic thinking or anticipation of negative outcomes.
  *Example:* A clear blue sky with a giant black storm cloud labeled "anxiety" approaching. Text: "That feeling something terrible is about to happen... all the time."
- **Excessive Worry** - Recurrent overthinking or exaggerated concerns about multiple life aspects.
  *Example:* A cartoon person spiraling in a thought cloud of disasters. Text: "Made a typo. Now convinced my boss will fire me."
- **Irritability** - Mood instability or frustration, often shown through abrupt reactions or sarcasm.
  *Example:* A figure steaming like a kettle with "rage mode" eyes. Text: "Someone breathing near me: exists — Me: explodes"

# 4 Methodology

This work proposes a novel multi-modal pipeline for classifying mental health memes-specifically, identifying symptoms of depression and anxiety-by integrating textual, visual, and figurative-comprehension components. Our approach enhances common sense grounding using image descriptions, meme text (OCR), and figurative prompts processed through a powerful large language model (Meta LLaMA 70B via NVIDIA NIM). The architecture is divided into two phases: Knowledge Fusion Construction and Retrieval Inference, as shown in the figure.

## 4.1 Text Extraction

We extract embedded meme text using EasyOCR, an open-source text extractor well-suited for noisy, real-world fonts and backgrounds. The resulting $O$i captures the meme's textual overlay, which often includes the core sarcastic or depressive message.

- Superior accuracy in meme text recognition compared to Tesseract, particularly for low-contrast or stylized fonts.

## 4.2 Visual Description Extraction

To capture latent visual semantics of memes (which often carry mental health cues via symbolism or dark humor), we utilize BLIP (Bootstrapped Language Image Pretraining) [Li et al., 2022] for visual captioning. BLIP generates a natural language description of the meme image. This augments the textual understanding, enabling our model to reason beyond OCR-visible text.

- Enables deeper semantic grounding of visual content compared to feature-only approaches like ResNet.
- Unlike [Mazhar et al., 2025], who omitted visual context, our inclusion of BLIP enriches figurative grounding with implicit visual elements.

## 4.3 Figurative Reasoning

Inspired by the GPT-based commonsense reasoning in [Mazhar et al., 2025], we employ Meta LLaMA 70B via NVIDIA NIM API to generate rich figurative explanations from a composite prompt comprising:

- BLIP-generated visual description*
- EasyOCR text*
- Cause-Effect
- Figurative Understanding
- Mental State

This prompt is sent to LLaMA to infer implicit meanings, metaphorical interpretations, or emotional tones.

- Our pipeline enhances the semantic depth by incorporating image-grounded prompts, unlike the text-only pipeline in the reference paper.

### 4.4 Knowledge Fusion

For each meme, the visual description, OCR text, and figurative reasoning output is embedded into a dense vector representation using Sentence-Transformer. These embeddings are then concatenated to form a unified knowledge representation vector $E$i for each meme. This allows meaningful retrieval.

The final concatenated embeddings [$o$i $\oplus$vi $\oplus$fi] (text + visual + figurative) are stored in the Knowledge Fusion DB for retrieval. This design ensures that reasoning, not just surface-level features, contributes to classification. This embedding represents the meme in a high-dimensional semantic space suitable for retrieval and classification.

$$\mathbf{K}_i = \mathbf{k}_i^{\text{CE}} \oplus \mathbf{k}_i^{\text{MS}} \oplus \mathbf{k}_i^{\text{FG}} \oplus \mathbf{k}_i^{\text{VD}}$$

where,
$K$i: Final knowledge embedding for meme i
$\oplus$: Concatenation of the vectors
Superscripts: CE = Cause-Effect, MS = Mental State, FG = Figurative, VD = Visual Description

$$\mathbf{K}_i = \mathbf{k}_i^{\text{Fig}} \oplus \left( \frac{1}{k} \sum_{j=1}^{k} \mathbf{e}_j^{\text{sim}} \right)$$

where,
$\mathbf{k}_i^{\text{Fig}}$: Figurative knowledge embedding (via Meta LLaMA 70B using OCR + visual description in prompt)
$\mathbf{e}_j^{\text{sim}}$: Embedding of $j$th similar meme retrieved using Sentence-BERT
k: Number of similar memes considered
Final vector: $K$i is used for classification

$$\begin{aligned}\mathcal{E} = &\ \Pi\left([\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n]\right)^{n \times d} \oplus \\ &\ \Pi\left([\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n]\right)^{n \times d} \oplus \\ &\ \Pi\left([\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]\right)^{n \times d}\end{aligned}$$

**where:**

- $\mathbf{o}_i$: Object-level tokens extracted from the meme image (e.g., object descriptions from BLIP).
- $\mathbf{r}_i$: Relation-level tokens such as sarcasm targets or OCR text.
- $\Pi(\cdot)$: Embedding function (e.g., Sentence-Transformer).
- $\oplus$: Concatenation operator.
- $n$: Number of token pairs (object-relation).
- $d$: Embedding dimension.

$$\mathbf{e}_k = (\Pi(\mathbf{o}_k) \oplus \Pi(\mathbf{r}_k)) \in R^{1 \times 3d}; \quad y = \arg\max_i \Phi(\mathbf{e}_k, \mathcal{E}_i) \tag{1}$$

**where:**

- $\mathbf{e}_k$: Joint representation of the $k$-th query meme.
- $\Phi(\cdot, \cdot)$: Similarity function (e.g., cosine similarity).
- $y$: Index of the most relevant explanation in $\mathcal{E}$.
- $\mathcal{E}_i$: $i$-th candidate explanation embedding from the fused matrix.

### 4.5 Retrieval, Training Classification

This stage retrieves the most semantically relevant knowledge for a meme and classifies it using a retrieval-augmented decoder. We perform a top-k similarity search (e.g., k=3 or 5) in the KF-DB using cosine distance over the $e$k vector. The top-k similar memes ([p, q, r]) provide supporting knowledge vectors [$E$p, $E$q, $E$r].

- Each new meme (during training or inference) retrieves k-nearest neighbor memes based on embedding similarity (cosine distance). Each retrieval brings relevant memes with shared figurative, emotional, or commonsense patterns.
- SentenceTransformer embeddings support cosine similarity-based retrieval.
- Robust to domain shifts due to multimodal fusion input.
- Encourages explanation-based retrieval and mitigates overfitting to surface-level features.
- This retrieval-guided reasoning mechanism differentiates our system from transformer-only baselines.

A transformer decoder is trained over this vector to predict mental health categories (e.g., anxiety labels). We use a multi-head decoder structure stacked in layers, enabling hierarchical attention to different aspects of input.

- Memory-augmented training leads to better generalization, especially with data sparsity.
- Concatenated retrieved embeddings act like external knowledge support.

### 4.6 Dataset-Specific Optimization

- For the Anxiety Dataset, the full FUSE-MH pipeline was applied, including figurative reasoning via Meta LLaMA, BLIP, and OCR, as this provided rich emotional context.
- For the Depression Dataset, we observed that EasyOCR + CLIP embeddings alone outperformed the full pipeline. Thus, for this dataset, CLIP was used directly on the meme image and OCR text, bypassing figurative reasoning.

### 4.7 Visual-Linguistic Embedding via CLIP for Depression Dataset

For the depression meme classification task, we adopted a distinct yet complementary approach within the broader FUSE-MH (Figurative, Fused, and Fine-Grained Mental Health) framework. Early experimentation revealed that CLIP (Contrastive Language–Image Pretraining) [Radford et al., 2021], when combined with meme-specific OCR text, was highly effective in capturing depressive cues - without requiring additional reasoning or retrieval-based knowledge infusion. Consequently, we designed a streamlined pipeline that prioritized performance and efficiency for this dataset.
Unlike the anxiety pipeline that relied on BLIP for visual description generation and LLM prompting for figurative reasoning, our depression-specific setup utilizes the CLIP model directly to encode the image-text interplay:

- OCR-based Caption Extraction: Text was extracted from meme images using EasyOCR, capturing explicit or subtly depressive language (e.g., “I’m broken but smiling”).
- Joint Embedding: The original meme image and OCR-extracted text were jointly embedded using CLIP’s ViT-B/32 model, which aligns visual and textual modalities into a shared semantic space. This strategy was sufficient to capture tone, affect, and symbolic alignment between text and image. This embedding $E$i represents the semantic fusion of both modalities and served as input for the downstream classification.

$$\mathbf{E}_i = \text{CLIP}(\text{image}_i, \text{text} = o_i)$$

- This CLIP-based variant thus balances semantic richness and computational efficiency, making it a practical and robust solution for depressive meme analysis — while still aligning with the philosophy of the broader FUSE-MH framework.

### 4.8 Prompts and Examples

To effectively ground meme understanding in both commonsense and figurative reasoning, our system formulates a structured query to the language model using the following enriched prompt:

**1. Cause-Effect:** Identify concrete causes or effects associated with the situation depicted in the meme.

**2. Figurative Interpretation:** Extract underlying metaphors, analogies, symbolic elements, or ironic tones that suggest deeper meanings.

**3. Mental State**: Describe the emotional or psychological states conveyed in the meme.

**4. Meme Text (via OCR):** Textual content overlaid on the meme image, often carrying figurative or emotionally expressive cues.

**5. Visual Description (via BLIP):** Automatically generated textual depiction of the image's visual scene, including entities, actions, and context.

**Visual description:** a girl with blonde hair and black shirt is sitting on a chair with a quote on it.

**OCR - Text:** Me as a child feeling that I was never heard. Me as an adult trying to help everyone.

**Figurative Reasoning:**

**Cause-Effect:** The meme suggests that speaker's childhood experience of feeling unheard or unhelped has led to stress issues.

**Figurative Understanding:** The meme uses a metaphorical representation of speaker's past and present to explain her mental state.

**Mental State:** The meme depicts a sense of empathy, compassion, and possibly anxiety in speaker as they strive to communicate to people.

## 5 Results

We evaluated our proposed framework FUSE-MH (Figurative-Understanding and semantics Embedding for Mental Health memes) on both the Anxiety and Depression datasets using standard classification metrics. For each dataset, we report the Macro F1 and Weighted F1 scores to account for class imbalance and multi-class nature of the task.

### 5.1 Depression Meme Classification

We observed that this minimalist yet semantically rich representation yielded the most effective performance for depression symptom recognition, without requiring additional figurative knowledge fusion. We report a Macro F1 score of 0.308 and a Weighted F1 score of 0.516 on the Depression dataset. The macro F1 score indicates the model's balanced performance across all classes, despite the skewed distribution. The weighted F1 score reflects the higher precision and recall for majority classes like Feeling Down and Low Self-Esteem.

### 5.2 Anxiety Meme Classification

The proposed FUSE-MH framework demonstrated superior performance on the anxiety meme classification task, effectively leveraging external figurative reasoning and retrieved commonsense knowledge to disambiguate nuanced mental health cues embedded in memes. We report a Macro F1 score of 0.56 and a Weighted F1 score of 0.58 on the Anxiety dataset.

The macro F1 score reflects strong balanced performance across all six anxiety symptom classes, including Restlessness, Worry Control, and Irritability, despite the inherent class imbalance. The weighted F1 score further indicates that the model maintains high precision and recall across both dominant and minority categories, benefiting from the retrieval-augmented training pipeline and the enriched figurative prompts generated by Meta LLaMA.

These results validate the effectiveness of our figurative interpretation strategy and highlight the potential of hybrid reasoning pipelines for complex understanding of affective memes.

| | | |
|---|---|---|
| Depression Macro F1 | CLIP + Transformer | 0.308 |
| Depression Weighted F1 | CLIP + Transformer | 0.516 |
| Anxiety Macro F1 | FUSE-MH | 0.568 |
| Anxiety Weighted F1 | FUSE-MH | 0.589 |

## 6 Discussion and Analysis

Our experimental results demonstrate the importance of modality-aware architecture design in mental health meme classification. Specifically, we observed that different modalities contribute differently across datasets:

- For the Depression dataset, CLIP-based multimodal embeddings (image + text) effectively captured the dominant visual and linguistic cues. The fusion of these modalities without additional figurative reasoning yielded better performance, indicating that depression-related memes often convey their message more directly through visual tone or text sentiment.

- In contrast, for the Anxiety dataset (detailed in later sections), memes exhibited more subtle, figurative expressions, requiring symbolic understanding and commonsense reasoning. This justified the need for our knowledge-infused framework (FUSE-MH) using Meta LLaMA 70B, BLIP, and EasyOCR to explicitly model mental state, cause-effect reasoning, and figurative cues.

Several observations emerge:

- Figurative language such as sarcasm and metaphor remains a significant challenge for baseline visual-language models. Our framework's incorporation of external commonsense reasoning improved classification for abstract symptoms like Impending Doom or Lack of Worry Control.

- Visual grounding through BLIP-generated captions proved crucial. When included in LLaMA prompts, these descriptions enhanced figurative interpretation by anchoring the symbolic meaning of the image.

- The asymmetric effectiveness of methods across datasets reinforces the idea that meme-based mental health expression is heterogeneous. Some symptom categories rely more on overt textual content, while others require deeper semantic and affective interpretation.

These insights underline the necessity of adaptive multimodal pipelines for different mental health conditions and symptom categories. Future

work can focus on class-specific modeling strategies, integration of additional emotion ontologies, and expansion to cross-cultural meme datasets.

## 7 Conclusion

In this work, we proposed a flexible, knowledge-infused multimodal framework—FUSE-MH to classify mental health-related memes, with a particular focus on anxiety and depression. Our experiments highlighted the complementary strengths of different components: while CLIP-based embeddings performed well for the depression dataset, anxiety classification required deeper reasoning, which was effectively addressed by incorporating figurative understanding and commonsense knowledge via Meta LLaMA 70B and retrieval-augmented prompts.

By leveraging visual description (BLIP), text extraction (EasyOCR), and semantic embedding fusion, our model was able to interpret the subtle emotional and symbolic cues embedded in memes. These results underscore the importance of modality-aware and knowledge-enriched approaches in handling complex, nuanced mental health content on social platforms.

## 8 Future Scope

- **Incorporating Larger Language Models:** While Meta LLaMA 70B provided strong figurative reasoning through prompt-based generation, future iterations could explore more advanced models like LLaMA 3-430B. These models offer improved contextual understanding and symbolic reasoning, which may better capture subtle affective cues and multi-layered figurative language common in mental health-related memes.

- **Direct Use of Visual Embeddings:** Currently, visual knowledge is infused indirectly through BLIP-generated descriptions. A potential enhancement lies in directly integrating visual embeddings (e.g., from CLIP-ViT or OpenCLIP) into the fusion pipeline. This could help the model capture fine-grained visual details such as expressions, posture, and background symbolism that textual descriptions may overlook.

Together, these extensions can push the boundary of commonsense and affective reasoning in multimodal mental health analysis, making the system more robust and scalable for real-world applications.

## References


[1] Machine Learning and Natural Language Processing in Mental Health: Systematic Review https://www.jmir.org/2021/5/e15708/

[2] Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques https://formative.jmir.org/2022/9/e36118/

[3] Assessing ML Classification Algorithms and NLP Techniques for Depression Detection: An Experimental Case Study https://arxiv.org/abs/2404.04284

[4] Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification https://arxiv.org/abs/2501.15321

[5] A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model https://onlinelibrary.wiley.com/doi/full/10.1155/2022/7893775

[6] Transformer-based language models for mental health issues: A survey https://www.sciencedirect.com/science/article/abs/pii/S0167865523000430

[7] Predicting Mental illness (Depression) with the help of NLP Transformers https://ieeexplore.ieee.org/abstract/document/10594036?casa_token=naayIURrdgMAAAAA:Oe2nXx3jwCswNo3mRvgRjjuNoyKikMtONNM8l8zMRCQvhO6rM-y96GAzDZ6iO4H2VwM0x9h