Mental Health Meme Classification
NLP Project Proposal
Group-12

Manan Aggarwal
2022273
manan22273@iiitd.ac.in

Shobhit Raj
2022482
shobhit22482@iiitd.ac.in

Souparno Ghose
2022506
souparno22506@iiitd.ac.in

Abstract

Mental health issues such as anxiety and depression have become increasingly prevalent on social media, with memes emerging as a prominent form of self-expression. In this work, we present a novel multimodal framework for the classification of mental health-related memes, tackling both single-label anxiety and multi-label depression classification tasks. Our approach leverages complementary information from textual and visual modalities using OCR-extracted text, mental health-specific language models (Mental-RoBERTa), and visual encodings from CLIP. We introduce a cross-attention fusion mechanism and a Mixture-of-Experts (MoE) gating module to effectively integrate representations, along with a contrastive loss that aligns modalities in a shared embedding space.

Our method demonstrates strong performance across both datasets. On the AxiOM Anxiety dataset, our model achieves a Macro-F1 score of 0.6851 and a Weighted-F1 of 0.6848, significantly outperforming the strongest baseline (OCR + Mental-BERT) by over 6%. On the RESTORE Depression dataset, our approach attains a Macro-F1 of 0.6606 and a Weighted-F1 of 0.6628, improving upon the best baseline by around 3%. Extensive ablation studies confirm the critical role of multimodal fusion and contrastive learning in enhancing model performance. These results validate the effectiveness of our architecture in modeling the nuanced semantics of mental health memes. The code for this project can be found in this GitHub Repository.

1 Introduction

In recent years, social media platforms have emerged as unconventional outlets for individuals to express their emotions, struggles, and personal experiences, often through humor and satire. Among these, internet memes have become a popular medium, especially among younger audiences, to communicate complex feelings in a lighthearted, digestible format. While seemingly humorous on the surface, many memes subtly reflect deeper emotional states such as anxiety, depression, or existential distress.

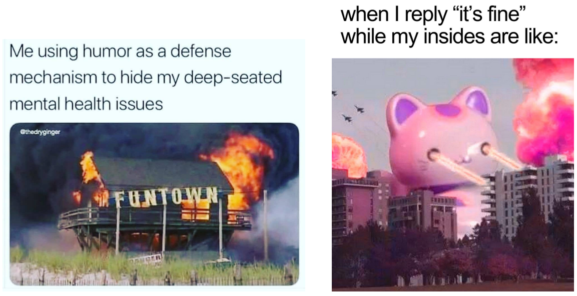
Me using humor as a defense mechanism to hide my deep-seated mental health issues
@thedryginger
FUNTOWN
when I reply "it's fine" while my insides are like:

Figure 1: Examples of memes reflecting underlying mental health issues.

Consider the two memes shown in Fig. 1. The first meme shows an amusement park building labeled "FUNTOWN" on fire, with the caption "Me using humor as a defense mechanism to hide my deep-seated mental health issues." This contrast between a fun-looking place and the destruction behind it conveys how people often use humor and sarcasm to hide their emotional vulnerability. The second meme shows a giant cartoon cat shooting laser beams from its eyes while everything around it is in chaos. The caption reads "when I reply 'it's fine' while my insides are like:". Even though it looks silly, the meme clearly represents the common experience of pretending everything is okay while feeling overwhelmed inside, a behavior often linked to anxiety and emotional suppression.

Despite the increasing prevalence of such content, automatic detection and categorization of mental health cues in memes remain underexplored in research. Most existing approaches to mental health detection rely solely on textual data such as tweets or Reddit posts, failing to account for the multimodal nature of memes, which combine text, imagery, and often figurative language. This multimodal blend can contain nuanced indicators of mental states that traditional unimodal systems overlook.

Our project aims to bridge this gap by developing a multimodal deep learning framework that can identify and classify mental health symptoms, specifically anxiety and depression from memes. By analyzing both textual and visual cues, including subtle or implied meanings, our system looks for hidden emotional signs that might otherwise go unnoticed. This work offers a step toward better understanding how mental health expressions manifest in multimodal content and explores how computational models can be used in interpreting them.

2 Related Work

The intersection of multimodal learning and mental health analysis has garnered significant attention, particularly in understanding how memes convey complex emotional states. Two pivotal studies have laid the groundwork in this domain:

Depression Symptom Classification (Yadav et al., 2023): introduced the RESTORE dataset, focusing on the fine-grained classification of depression symptoms depicted in memes. Each meme in this dataset is annotated with one or more of eight depression symptoms derived from the PHQ-9 questionnaire. Their approach emphasizes the integration of textual and visual modalities, employing transformer-based architectures to capture the nuanced interplay between meme text and imagery. Notably, they implemented orthogonal constraints to ensure non-redundant feature learning across modalities, enhancing the model's ability to discern subtle depressive cues.

Anxiety Symptom Classification (Mazhar et al., 2025): Building upon this foundation, developed the AxiOM dataset, targeting the classification of anxiety symptoms in memes. Recognizing the limitations of existing models in interpreting figurative language and commonsense knowledge, they proposed the M3H framework. This framework enriches multimodal language models by infusing them with domain-specific knowledge and figurative reasoning capabilities. Through comprehensive evaluations, M3H demonstrated superior performance in capturing the intricate expressions of anxiety within memes, highlighting the importance of integrating external knowledge sources for nuanced understanding.

Our work builds upon these studies by exploring modular baselines and fusion strategies across both datasets. While prior research has focused on specific end-to-end fusion architectures, we analyze the incremental contribution of each modality OCR text, visual features, and LLM-generated reasoning using transformer-based encoders. Furthermore, we introduce a lightweight multimodal attention mechanism coupled with a contrastive learning objective, that encourages the fused representations to stay semantically close to both textual and visual embeddings, thereby improving the coherence and robustness of the learned features.

3 Methodology

Our proposed approach integrates multimodal and figurative information from memes to robustly classify mental health-related content. The overall pipeline consists of several key modules, each designed to capture a distinct aspect of the input data. In our experiments, we address two tasks: (i) single-label anxiety classification using the AxiOM dataset and (ii) multi-class depression classification using the RESTORE dataset. We combine the classification loss with a contrastive loss inspired by SimCLR (Chen et al., 2020) to better align representations across modalities for both the models.

The input to the model is the concatenation of OCR-extracted text and figurative reasoning output from LLaVA, along with the meme image. Let:

$$X_{\text{OCR}} \in \mathbb{R}^{L_{\text{ocr}}}, \quad X_{\text{LLAVA}} \in \mathbb{R}^{L_{\text{llava}}},$$

be the tokenized OCR text and figurative reasoning text, respectively. We define the concatenated text input as:

$$X = X_{\text{OCR}} + \texttt{[SEP]} + X_{\text{LLAVA}} \in \mathbb{R}^{L},$$

where $L = L_{\text{ocr}} + L_{\text{llava}}$.

3.1 Pipeline Overview

1. **OCR and Pre-processing:** We first extract textual content from meme images using Google Docs OCR via Google App Script. To ensure compatibility and fair evaluation with prior work, we align our dataset by removing specific classes and categories that were excluded in the referenced studies.

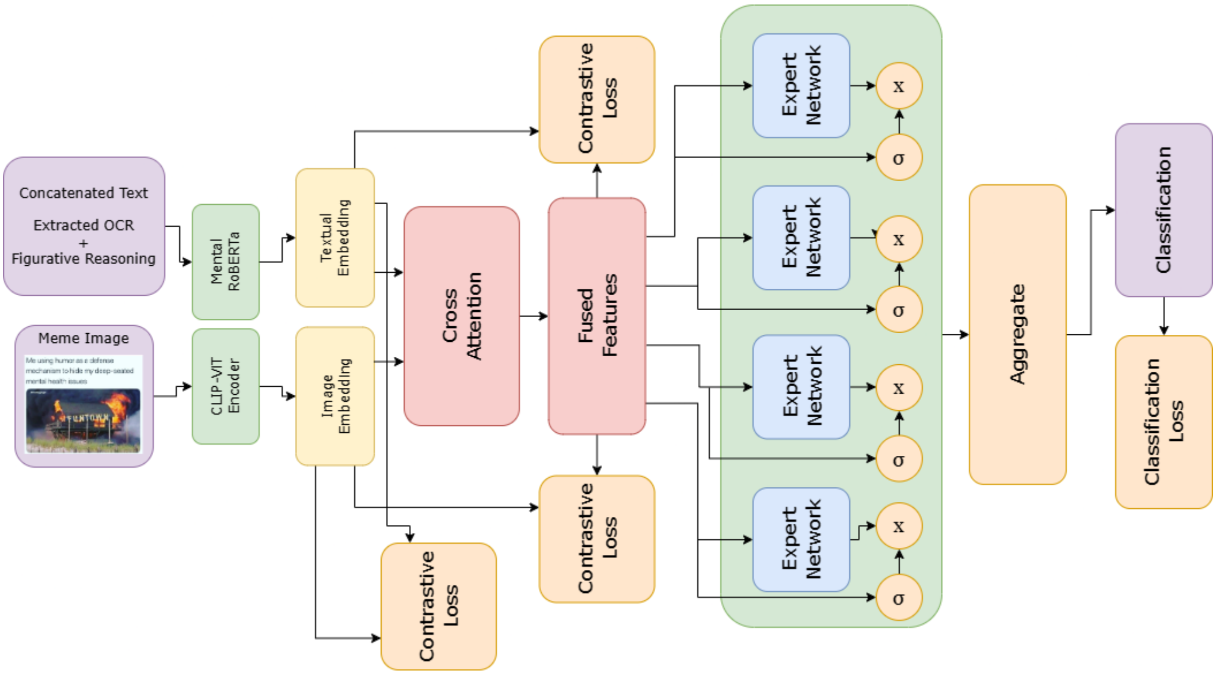

Figure 2: Our Complete Pipeline

2. **Figurative Reasoning:** To capture the implicit, non-literal cues present in memes, we use a large language model (LLM), `LLaVA` (Liu et al., 2023) to generate textual reasoning about the image context. Let this output be represented as X_{LLAVA}. It enriches the original OCR text by providing additional figurative context.

3. **Textual Representation:** The concatenated text X is fed into a `Mental-RoBERTa Transformer` (Ji et al., 2022) to obtain rich, context-aware embeddings:

$$T = f_{\text{MR}}(X) \in \mathbb{R}^d,$$

where d is the model's hidden size. This representation captures semantic nuances essential for understanding emotional content.

4. **Visual Embedding:** In parallel, images are processed using a `CLIP-based Encoder` (Wang et al., 2022) (Vision Transformer) to extract deep visual features. Let the input image be $I \in \mathbb{R}^{H \times W \times 3}$, and the encoder output be:

$$V = f_{\text{CLIP}}(I) \in \mathbb{R}^d,$$

which encodes spatial and structural information correlating with emotional cues.

5. **Multimodal Fusion via Cross-Attention:** To integrate the textual and visual modalities, we first project the embeddings into a common fusion space of dimension d_f:

$$T' = W_T T, \quad V' = W_V V,$$

$$W_T, W_V \in \mathbb{R}^{d_f \times d}$$

We then employ a custom **Cross-Attention Layer** that fuses these projected features. For each cross-attention block, we define:

$$Q = W_q T', \quad K = W_k V', \quad V_v = W_v V',$$

where $W_q, W_k, W_v \in \mathbb{R}^{d_f \times d_f}$. The attention scores are computed as:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right),$$

with d_h being the attention head size. The fused feature from the block is:

$$F = \text{LayerNorm}(T' + W_O(A \cdot V_v))$$

The mean of the outputs of multiple cross-attention blocks is taken to obtain the combined multimodal feature.

6. **Expert Fusion and Mixture-of-Experts (MoE):** The fused feature is then processed by 4 expert networks $E_i(\cdot)$, $i = 1, \ldots, 4$ and

combined via an MoE gating mechanism, inspired by the sparse MoE framework (Shazeer et al., 2017). If F denotes the combined feature, the expert outputs are:

$$E_i = \text{Expert}_i(F) \quad \text{for } i = 1, \ldots, 4,$$

and the final fused output is:

$$F_{\text{MoE}} = \sum_{i=1}^{4} g_i(F) \cdot E_i,$$

where $g_i(F)$ are the gating weights obtained by applying a softmax over the MoE gate's output.

7. **Final Classification:** The MoE output F_{MoE} is then passed to a `Feed-Forward-based Classifier` which outputs the class logits:

$$\hat{y} = f_{\text{CLS}}(F_{\text{MoE}}) \in \mathbb{R}^C,$$

where C denotes the number of classes. For the `AxiOM` dataset (anxiety classification), the model is optimized using a combination of cross-entropy loss and a contrastive loss component. For the `RESTORE` dataset (depression classification), a standard binary cross-entropy loss is applied over all the classes with contrastive loss component.

3.2 Loss Functions and Training Strategy

We optimize our multimodal model by minimizing a joint loss that combines a primary classification loss with a contrastive loss that aligns the multimodal representations. The overall loss is defined as:

$$\mathcal{L} = (1 - \lambda_{\text{cont}})\, \mathcal{L}_{\text{cls}} + \lambda_{\text{cont}}\, \mathcal{L}_{\text{contrastive}},$$

where $\lambda_{\text{cont}} \in [0, 1]$ is a hyperparameter that controls the relative contribution of the contrastive loss.

Contrastive Loss: Inspired by the SimCLR framework (Chen et al., 2020), the contrastive loss operates on pairs of normalized embeddings. Given two sets of embeddings from different modalities—denoted $z^{(1)}$ and $z^{(2)}$ (e.g., the multimodal embedding versus the text or image embedding) we first concatenate them:

$$Z = \text{concat}\left(z^{(1)}, z^{(2)}\right) \in \mathbb{R}^{2N \times d},$$

where N is the batch size and d is the embedding dimension. We then compute the similarity matrix:

$$S = \frac{ZZ^\top}{\tau} \in \mathbb{R}^{2N \times 2N},$$

with the temperature τ scaling the similarities. A binary mask M is applied to zero out self-similarities:

$$M_{ij} = \begin{cases} 0 & \text{if } i = j, \\ 1 & \text{otherwise.} \end{cases}$$

Then the loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2N} \sum_{i=1}^{2N} \ell(i),$$

where $\ell(i)$ is the cross-entropy loss that treats the i-th row of the similarity matrix (with masked diagonal) as logits and the target label is defined by pairing the two corresponding modalities.

Anxiety Classification (AxiOM): For the single-label anxiety task, the classification loss $\mathcal{L}_{\text{cls}}$ is defined using the standard cross-entropy loss. Let the predicted logits be $\hat{y} \in \mathbb{R}^C$ and the ground-truth label $y \in \{1, \ldots, C\}$. Then,

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(\hat{y}_y)}{\sum_{j=1}^{C} \exp(\hat{y}_j)}.$$

Thus, the final loss for anxiety is:

$$\mathcal{L}^{(\text{anxiety})} = (1 - \lambda_{\text{cont}})\, \mathcal{L}_{\text{CE}} + \lambda_{\text{cont}}\, \mathcal{L}_{\text{contrastive}}$$

Depression Classification (RESTORE): For the multi-label depression task, the classification loss is computed by applying binary cross-entropy independently to each of the C labels. For a given sample with ground-truth label vector $y \in \{0, 1\}^C$ and the predicted probabilities $\hat{y} \in [0, 1]^C$ (obtained by applying sigmoid to the logits), the loss is:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{C} \sum_{c=1}^{C} \Big(y_c \log(\hat{y}_c) + (1 - y_c) \log(1 - \hat{y}_c) \Big).$$

Then, the total loss for depression is given by:

$$\mathcal{L}^{(\text{depression})} = (1 - \lambda_{\text{cont}})\, \mathcal{L}_{\text{BCE}} + \lambda_{\text{cont}}\, \mathcal{L}_{\text{contrastive}}$$

Training Strategy: We partition the training data into training and validation subsets. During each epoch, we compute contrastive losses between `Multimodal and text embeddings`, `Multimodal and image embeddings` and `Text and image embeddings`, which is then averaged out to obtain $\mathcal{L}_{\text{contrastive}}$ and then the appropriate loss according to the problem is computed to obtain the final loss

$\mathcal{L}$. Performance metrics-specifically, `macro-F1` and `weighted-F1` on both the training and validation sets are also computed to ensure that our model generalizes well and adequately handles class imbalances. The best-performing model on the validation set is selected for final evaluation on the test set.

4 Dataset, Experimental Setup, and Results/Findings

4.1 Datasets

AxiOM Anxiety Dataset The AxiOM dataset consists of memes annotated with one anxiety category. The class distributions for the training, validation, and test splits are given in Table 1. In total, the training set includes 2,153 samples, the validation set 307 samples, and the test set 615 samples.

Anxiety Category	Train	Validation	Test
Nervousness	373	53	106
Lack of Worry Control	331	47	94
Excessive Worry	322	46	92
Difficulty Relaxing	356	51	102
Restlessness	405	58	116
Impending Doom	366	52	105
Total	2,153	307	615

Table 1: Class Distribution for AxiOM Anxiety Dataset

RESTORE Depression Dataset The RESTORE dataset is used for multi-label depression symptom classification. Each meme may exhibit one or more depression symptoms. Table 2 details the class distributions. Overall, the training set contains 9,023 samples, the validation set 522 samples, and the test set 721 samples.

Depression Symptom	Train	Validation	Test
Lack of Interest	471	45	71
Feeling Down	2,085	195	218
Eating Disorder	1,939	49	92
Sleeping Disorder	1,562	45	79
Low Self-Esteem	855	85	114
Concentration Problem	595	42	66
Self-Harm	1,516	61	81
Total	9,023	522	721

Table 2: Class Distribution for RESTORE Depression Dataset

4.2 Experimental Setup

All experiments were conducted on a Kaggle Notebook environment utilizing 2 GPUs. We first acquired figurative reasoning using the `LLAVA` model, and then used this output in addition to OCR-extracted text and image features as input to our model.

We implemented three baseline models:

1. **OCR + BERT (Devlin et al., 2019):** Fine-tuned BERT using OCR text only.
2. **OCR + Mental-BERT (Ji et al., 2022):** Fine-tuned Mental-BERT using OCR text.
3. **OCR + LLAVA Reasoning + Mental-BERT:** Combined OCR text and LLAVA-generated figurative reasoning as input to Mental-BERT.

For all experiments, we used a consistent set of hyperparameters optimized for stable training and performance. The contrastive loss coefficient λ_{cont} was set to 0.3 to balance its influence with the classification loss. The fusion dimension, representing the shared space for cross-modal interaction, was set to 1024. We used a maximum sequence length of 512 tokens to accommodate longer textual inputs, with a batch size of 16 and trained all models for 30 epochs. The learning rate was set to $2e-5$, following common fine-tuning practices for transformer-based models. For contrastive learning, the temperature parameter was adopted from the (Chen et al., 2020) paper to maintain effective separation in the embedding space.

4.3 Results and Findings

Tables 3 and 4 summarize the test metrics (macro-F1 and weighted-F1) for the baseline models and our final approach.

AxiOM Anxiety Results:

Model	Macro-F1	Weighted-F1
OCR + BERT	0.6163	0.6143
OCR + Mental-BERT	0.6235	0.6232
OCR + LLAVA + Mental-BERT	0.6183	0.6173
Proposed Approach	**0.6851**	**0.6848**

Table 3: Test Metrics for AxiOM Anxiety Dataset Baselines and Proposed Approach

RESTORE Depression Results:

Model	Macro-F1	Weighted-F1
OCR + BERT	0.6355	0.6347
OCR + Mental-BERT	0.6313	0.6249
OCR + LLAVA + Mental-BERT	0.6298	0.6263
Proposed Approach	**0.6606**	**0.6628**

Table 4: Test Metrics for RESTORE Depression Dataset Baselines and Proposed Approach

5 Discussion/Analysis/Observations

5.1 Performance Comparison

To evaluate the effectiveness of our proposed approach, we compare its performance against a set of strong baselines on both the AxiOM Anxiety and RESTORE Depression datasets. The baseline models include combinations of OCR-extracted text features with BERT and Mental-BERT, as well as a multimodal variant incorporating image features via LLAVA alongside Mental-BERT for text.

For the AxiOM Anxiety dataset, our proposed model achieves a significant improvement with a Macro-F1 score of 0.6851 and a Weighted-F1 score of 0.6848. In comparison, the best-performing baseline (OCR + Mental-BERT) reaches a Macro-F1 of 0.6235 and a Weighted-F1 of 0.6232. This demonstrates a relative gain of over 6 percentage points in both metrics, highlighting the advantage of our multimodal fusion, cross-attention mechanism, and contrastive alignment.

On the RESTORE Depression dataset, our approach similarly outperforms all baselines, achieving a Macro-F1 of 0.6606 and a Weighted-F1 of 0.6628. The closest baseline (OCR + BERT) achieves 0.6355 Macro-F1 and 0.6347 Weighted-F1, showing that our model delivers consistent improvements in identifying depression categories as well.

These results clearly establish the effectiveness of our proposed method in capturing both visual and textual cues from mental health-related memes, and in modeling the intricate relationships between modalities for accurate classification.

5.2 Ablation Study

To better understand the contribution of each component in our multimodal framework, we conducted a series of ablation experiments on both the Anxiety (`AxiOM`) and Depression (`RESTORE`) tasks. In our experiments, we systematically removed or replaced individual modules, including:

1. **Contrastive Learning Removal:** Excluded the contrastive loss to evaluate its impact on aligning representations across modalities.

2. **Fusion Module Replacement:** Replaced the cross-attention fusion mechanism with a simple concatenation of features.

3. **Unimodal (No Image Encoding):** Evaluated performance using only textual features (OCR-extracted text with figurative reasoning).

4. **Unimodal (No Text Encoding):** Evaluated performance using only visual features (removing OCR and figurative reasoning).

The results of these experiments are summarized in Table 5.

The ablation study reveals several important insights. First, the complete multimodal model achieves the best performance for both tasks, with a Macro-F1 score of 0.6851 (Anxiety) and 0.6606 (Depression). Removing the contrastive learning component consistently leads to a decrease in performance, highlighting its role in aligning features across modalities. Additionally, replacing the cross-attention fusion mechanism with simple concatenation also results in inferior performance, underscoring the effectiveness of our dynamic fusion strategy.

A particularly notable observation is the significant drop in performance when text encoding is removed, indicating that OCR-extracted text (supplemented by figurative reasoning) is a critical source of information for both anxiety and depression classification. In contrast, the removal of image encoding causes a moderate decrease in performance, suggesting that while visual cues contribute meaningfully, they are secondary to the textual modality in these tasks.

Overall, these results validate the importance of each module in our pipeline, especially the fusion mechanism and contrastive learning, which together enhance the model's ability to capture subtle multimodal signals indicative of mental health states.

Configuration				Depression		Anxiety	
CL	Fusion	Img	Text	Macro-F1	Weighted-F1	Macro-F1	Weighted-F1
✗	✓	✓	✓	0.6454	0.6379	0.6469	0.6471
✓	✗	✓	✓	0.6504	0.6354	0.6503	0.6487
✗	✗	✗	✓	0.6445	0.6527	0.6606	0.6590
✗	✗	✓	✗	0.4187	0.4307	0.4845	0.4852
✓	✓	✓	✓	0.6606	0.6628	0.6851	0.6848

Table 5: Experimental results for various model configurations. A tick (✓) indicates the component is used, and a cross (✗) indicates it is removed.

5.3 Error Analysis

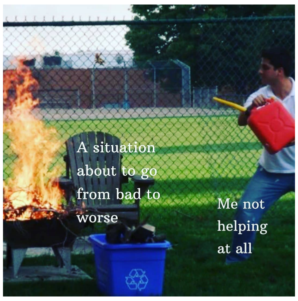

Figure 3: Anxiety Dataset (TE-573); Predicted: Impending Doom; Actual: Lack of Worry Control
OCR Text: **A situation about to go from bad to worse Me not helping at all**

Fig. 3 Shows that the model predicted a sample from the `Lack of Worry Control` class as `Impending Doom`, showcasing that even though the model is aware of worsening mental conditions and how they should be categorized, it is yet unable to correctly determine the emphasis of the kind of mental condition being targeted with the model overemphasizing on worry control to call it doom.

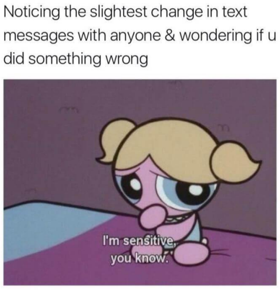

Figure 4: Anxiety Dataset (TE-213) Predicted: Excessive Worry, Actual: Nervousness
OCR Text: **Noticing the slightest change in text messages with anyone & wondering if u did something wrong I'm sensitive, you know.**

Fig. 4 illustrates a misclassification in which the model predicted `Excessive Worry` for a sample that should be labeled as `Nervousness`. The OCR text and figurative cues suggest that the subject is reacting sensitively to minor social cues. However, the model overestimates the level of anxiety by assigning an overly strong label, indicating that its feature extraction is emphasizing sensitivity more dramatically than intended.

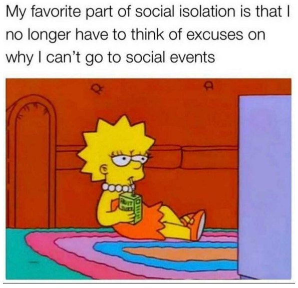

Figure 5: Depression Dataset (TE-146) Predicted: Feeling Down, Actual: Lack of Interest
OCR Text: **My favorite part of social isolation is that I no longer have to think of excuses on why I can't go to social events**

Fig. 5 shows a case where the model labeled the sample as `Feeling Down` rather than accurately capturing the underlying `Lack of Interest`. The OCR and figurative reasoning point to social disengagement and diminished motivation, a subtler state than general sadness. This suggests the model might be conflating low energy with overt sadness instead of recognizing the indifference or apathy inherent in lack of interest.

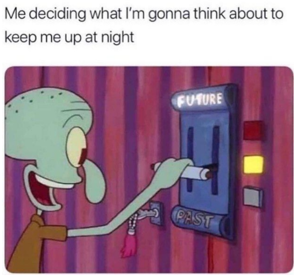

Figure 6: Wrong Prediction For Depression Dataset (TE-520) Predicted: Sleeping Disorder; Actual: Eating Disorder, Low Self-Esteem
OCR Text: **Me deciding what I'm gonna think about to keep me up at night FUTURE PAST**

In Fig. 6, the model predicts `Sleeping Disorder` even though the true labels indicate an underlying issue with `Eating Disorder` and `Low Self-Esteem`. The OCR and the figurative cues, such as the whimsical interaction with a "Future" button, imply deeper psychological struggles related to self-image and possibly disordered eating behaviors, rather than a mere sleep problem. This misclassification points to a potential need for the model to better differentiate between various internal distress signals.

6 Conclusion and Future Work

6.1 Conclusion

In this project, we presented a novel multimodal framework for the classification of mental health memes that leverages textual, visual, and figurative reasoning cues. By integrating OCR-extracted text with LLM-generated figurative reasoning and deep visual features through a cross-attention fusion mechanism and Mixture-of-Experts module, our model effectively captures the complex semantic nuances embedded within memes. Experimental results on the AxiOM Anxiety and RESTORE Depression datasets demonstrate that our approach significantly outperforms standard baselines, with improvements of over 6% in Macro-F1 for anxiety and notable gains for depression classification as well. Extensive ablation studies further confirm the critical roles of multimodal fusion and contrastive learning in enhancing overall performance.

6.2 Future Work

Looking ahead, there are several promising avenues for future work. One area of focus is the exploration of more advanced fusion techniques, such as dynamic gating based on adaptive modality weighting or deeper integration of contextual and semantic knowledge through graph-based models. Finally, further research into explainability methods for multimodal models would provide greater transparency on how various cues contribute to classification decisions, thereby fostering trust and practical deployment in real-world mental health support systems.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

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. *Preprint*, arXiv:2501.15321.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Preprint*, arXiv:1701.06538.

Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. 2022. Learning transferable human-object interaction detector with natural language supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 929–938.

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)*, pages 8890–8905, Toronto, Canada. Association for Computational Linguistics.