Mental Health Meme Classification

Aaditya Bhargav
IIIT Delhi

Shashank Pathak
IIIT Delhi

Anand Kumar
IIIT Delhi

Abstract

This proposal presents a novel multimodal framework to enhance the classification of mental health memes by integrating figurative reasoning and commonsense knowledge into visual and textual models. Our approach is designed to overcome limitations of current models in interpreting subtle, figurative expressions used in memes that convey mental health symptoms.

1 Introduction

In the digital age, memes have become one of the most popular modes of communication, especially among younger generations. These image-based artifacts often blend humor, sarcasm, irony, and visual storytelling to express complex emotional states. Increasingly, memes are being used to communicate experiences related to mental health—particularly depression and anxiety. This growing trend is not only visible across platforms like Reddit, Instagram, and Twitter but is also reflected in the emergence of entire communities devoted to sharing such content.

Despite their apparent humor, mental health-related memes often carry subtle but powerful emotional undertones. For example, a meme might depict a cartoon character saying "I'm okay" while surrounded by flames—an ironic portrayal of internal distress masked by outward indifference. In another, a relatable image may be accompanied by text like "Getting out of bed today was my biggest accomplishment," hinting at symptoms such as fatigue and loss of motivation. While humans can often intuitively recognize these underlying sentiments, automatic systems struggle to decode the figurative language, context, and visual cues embedded within.

Conventional mental health monitoring systems primarily rely on text-only data such as journal entries, tweets, or chat transcripts. However, they fall short when applied to memes, which require understanding both textual content (often humor-laden and ambiguous) and visual elements (e.g., facial expressions, setting, or body language). Furthermore, mental health expressions in memes often involve multi-label categorizations—for example, a single meme could simultaneously express hopelessness, social withdrawal, and irritability.

This project aims to bridge this gap by designing a robust multimodal framework that can accurately classify meme content into mental health symptom categories. Specifically, we focus on:

- **Multi-label classification of depression symptoms**—where each meme may belong to multiple depressive subcategories such as Anhedonia, Fatigue, or Worthlessness.
- **Single-label classification of anxiety symptoms**—where each meme is tagged with one dominant anxiety subcategory such as Excessive Worry or Restlessness.

By leveraging both visual and textual modalities, and enriching the model with commonsense reasoning and figurative understanding, this project contributes toward building empathetic AI systems capable of decoding nuanced human emotions from informal, user-generated content like memes. The ultimate goal is to support early detection and awareness of mental health conditions through passive analysis of online media.

2 Related Work

Prior studies have explored text-based mental health detection using social media posts (e.g., Reddit, Twitter). Studies like Cohan et al. (2018) used LSTM and transformer-based models to identify depressive language. On the visual side, recent work has utilized convolutional neural networks (CNNs) to detect emotional signals from images. The emergence of multimodal models, such as CLIP and

VisualBERT, has improved performance in tasks requiring joint image-text understanding. However, few approaches address the unique, figurative, and humorous nature of memes. The paper we reference introduces a novel formulation of this problem by including commonsense reasoning and visual-text fusion for better symptom prediction.

3 Methodology

We design a multimodal classification pipeline tailored for detecting depression and anxiety symptoms from internet memes. Our approach integrates both visual and textual information by leveraging state-of-the-art transformer models and OCR-based preprocessing.

- **OCR Extraction:** For each meme image, we use Tesseract OCR to extract textual content embedded within the image (e.g., captions or dialogues). This step is essential, as the majority of memes convey their context or humor through text superimposed on visuals.

- **Text Embedding:** We experiment with multiple text encoders to process the OCR-extracted content:
  - **Sentence-BERT (SBERT):** Used in the CLIP-only model for fast sentence-level embedding.
  - **MentalBERT:** A BERT variant pretrained on mental health forums, used to encode domain-specific language more effectively.
  - **RoBERTa:** Used as an alternative transformer encoder for robust semantic understanding.

- **Visual Embedding:** All models use the CLIP ViT-B/32 visual encoder to convert meme images into semantically meaningful embeddings. The pretrained CLIP model captures both low-level visual features and high-level contextual representations.

- **Fusion:** We implement an early fusion strategy to combine the visual and textual representations. Specifically:
  - Each text encoder produces a 768-dimensional embedding (for BERT variants).
  - CLIP produces a 512-dimensional image embedding.
  - These embeddings are concatenated to form a fused 1280-dimensional (or 896-dimensional for SBERT) representation, which is fed into a classifier.

- **Figurative Reasoning Module:** Although in its preliminary stage, we include a placeholder figurative reasoning component that enriches the OCR text with hypothetical cause-effect or sarcastic context (to be extended in future iterations with GPT-like reasoning).

- **Classification Heads:** We design task-specific classifier heads:
  - For the **depression dataset** (multi-label classification), we use a linear layer with `sigmoid` activation to allow prediction of multiple co-occurring depression symptoms.
  - For the **anxiety dataset** (single-label classification), we use a linear layer with `softmax` activation to predict exactly one anxiety category per meme.

- **Loss Functions:**
  - For depression classification, we use **Binary Cross-Entropy Loss** (`BCEWithLogitsLoss`) to handle multi-label output.
  - For anxiety classification, we use **Cross-Entropy Loss**, suitable for single-label multi-class problems.

- **Implemented Models Overview:** We evaluate the performance of different fusion models separately on the depression and anxiety datasets:

  Depression Dataset (Multi-label classification):
  - **M3H Model:** A baseline multimodal model that uses OCR text, CLIP embeddings, and a lightweight classifier.
  - **CLIP + Sentence Transformer:** Uses Sentence-BERT for text encoding and CLIP for visual encoding. Embeddings are concatenated and passed through a linear classifier.
  - **CLIP + MentalBERT:** Text is encoded using MentalBERT, which captures mental health–specific textual cues more effectively. Combined with CLIP visual embeddings.
  - **CLIP + RoBERTa:** RoBERTa serves as a robust general-purpose encoder, fused with CLIP visual features.

Anxiety Dataset (Single-label classification):

  - **CLIP + Sentence Transformer:** Utilizes SBERT for textual features and CLIP for image representation. Classifier head is optimized for categorical output.
  - **CLIP + MentalBERT:** Leverages mental health–specific transformer embeddings to improve anxiety class separation.
  - **CLIP + RoBERTa:** Applies RoBERTa's deep semantic encoding for the extracted meme text and fuses it with CLIP image embeddings.

- **Training Protocol:** Models are trained using the Adam optimizer with a learning rate of 5×10^{-5} for 10 epochs. We use batch sizes of 4 for both depression and anxiety datasets, and evaluate model performance on training data using accuracy, macro-F1, and weighted-F1 metrics.

4 Dataset, Experimental Setup, and Results

4.1 Dataset Description

We utilize two curated datasets for the task of meme-based mental health classification, focusing on symptoms of depression and anxiety:

- **Depression Dataset (Multi-Label):** This dataset contains meme images labeled with one or more depression-related symptoms. Each sample can belong to multiple classes simultaneously, such as Hopelessness, Anhedonia, Fatigue, and Social Withdrawal. The multi-label format makes this task more complex, as models must capture co-occurrence patterns and subtle semantic overlaps.

- **Anxiety Dataset (Single-Label):** Each meme is annotated with exactly one anxiety-related symptom from a fixed set of categories such as Excessive Worry, Nervousness, Restlessness, or Difficulty Relaxing. This dataset requires single-label classification, which is comparatively more straightforward than the depression task.

Each meme in both datasets contains embedded textual content, often containing sarcasm, humor, or figurative language. We use OCR to extract this text, which is then combined with the visual features to form a multimodal input.

4.2 Experimental Setup

All models were implemented using PyTorch and trained on NVIDIA Tesla V100 GPUs. The following configurations were used across all experiments:

- Learning rate: 5×10^{-5}
- Optimizer: Adam
- Batch size: 4
- Number of epochs: 10
- Loss functions:
  - Binary Cross-Entropy Loss for the depression (multi-label) task
  - Cross-Entropy Loss for the anxiety (single-label) task

Three multimodal architectures were evaluated across both datasets:

1. **CLIP + Sentence-BERT:** Uses CLIP for image features and Sentence-BERT for text embeddings. Embeddings are concatenated before being passed to the classifier.

2. **CLIP + MentalBERT:** Combines CLIP with MentalBERT, which is pretrained on mental health-related text. This model aims to capture domain-specific nuances.

3. **CLIP + RoBERTa:** Uses RoBERTa for robust text encoding, aiming to generalize better across diverse meme captions.

4.3 Results and Findings

4.3.1 Anxiety Dataset (Single-Label Classification)

The best performance on the anxiety dataset was achieved by the CLIP + Sentence-BERT model, with an accuracy of 56.81%, a macro-F1 score of 52.76%, and a weighted-F1 score of 55.48%. This indicates its strong generalization to short, informal meme texts.

The CLIP + MentalBERT model followed with an accuracy of 53.67%, a macro-F1 of 48.35%, and a weighted-F1 of 51.74%. This suggests that domain-specific text representations are helpful but not sufficient to outperform general-purpose models in simpler tasks.

CLIP + RoBERTa significantly underperformed with an accuracy of 38.81%, a macro-F1 of 31.87%,

and a weighted-F1 of 35.30%. This may be due to the model's sensitivity to informal language and difficulties adapting to noisy, sarcastic meme captions.

4.3.2 Depression Dataset (Multi-Label Classification)

On the depression dataset, which is inherently more challenging due to its multi-label nature, the CLIP + RoBERTa model performed the best with an accuracy of 48.67%, a micro-F1 score of 45.42, and a weighted-F1 score of 46.12. This indicates that RoBERTa is capable of modeling multiple co-occurring symptom types effectively when paired with visual context.

Interestingly, the CLIP + MentalBERT model achieved a slightly higher raw accuracy of 49.21% but fell behind in micro-F1 (43.42) and weighted-F1 (41.22). This suggests that while it captures individual labels well, it struggles with label co-occurrence and balance.

The CLIP + Sentence-BERT model performed poorly in this setting, with an accuracy of just 10.58%, a micro-F1 score of 1.77, and a weighted-F1 score of 7.76. The results highlight the limitations of simple sentence embeddings in multi-label, nuanced meme analysis.

4.3.3 Observations

- Models performed better on the anxiety dataset than the depression dataset, likely due to the simpler single-label structure.
- Sentence-BERT is lightweight and effective for single-label classification but fails to capture complex semantic relationships in multi-label settings.
- RoBERTa excels in depression classification, benefiting from its depth and attention mechanisms to understand nuanced emotional content across multiple categories.
- MentalBERT performs reasonably well in both settings, indicating the advantage of domain-specific pretraining, though it requires further tuning to handle multi-label dependencies.

4.3.4 Observations

- Models trained on the anxiety dataset achieved higher F1 scores than those on depression, likely due to the simpler, single-label nature of the task.
- RoBERTa performs well in multi-label classification (depression) but fails in anxiety classification, suggesting that task structure influences which text encoder performs best.
- Sentence-BERT, while efficient, lacks the depth required for nuanced multi-label learning and figurative interpretation.
- The results also highlight the importance of domain-specific pretraining, as seen in MentalBERT's competitive performance across both tasks.

4.4 Experimental Setup

Models were trained on NVIDIA Tesla V100 GPUs using PyTorch. We used Adam optimizer with a learning rate of 2×10^{-5}. Batch size was set to 32, and training was conducted for 10 epochs.

4.5 Results

Our system achieved the following:

- **Depression (multi-label):** Macro-F1: 0.48, Weighted-F1: 0.51
- **Anxiety (single-label):** Macro-F1: 0.53, Weighted-F1: 0.56

5 Discussion and Observations

Our experiments highlight several key insights into the challenges and behavior of multimodal models for mental health meme classification. The discussion below focuses on performance differences, model behaviors across tasks, and implications for future improvements.

5.1 Task Complexity

We observed a clear distinction in difficulty between the two tasks. The anxiety classification task, being a single-label classification problem, is relatively straightforward. Most models performed well, especially when text embeddings were generated using lightweight yet semantically strong encoders such as Sentence-BERT. In contrast, the depression classification task involves multi-label outputs, requiring the model to simultaneously detect co-occurring mental health symptoms. This adds significant complexity, as models must learn to disentangle overlapping representations and understand subtle contextual cues that may indicate multiple conditions.

5.2 Performance Across Models

- **Sentence-BERT:** While Sentence-BERT performed the best on the anxiety dataset, it struggled heavily on the depression dataset. This suggests that while it is capable of capturing semantic representations in short, literal texts, it lacks the contextual depth required to process ambiguous and figurative content found in multi-label memes.

- **MentalBERT:** The domain-specific MentalBERT model showed moderate performance on both tasks. Although it was designed to understand mental health–related text, its representations might be overly tuned to long-form and clinical language (e.g., Reddit or therapy transcripts), which may not align well with short, humorous, and sarcastic meme text. However, its competitive scores suggest strong potential for improvement with fine-tuning on meme-specific data.

- **RoBERTa:** RoBERTa performed exceptionally well on the depression dataset, likely due to its deep attention mechanisms and ability to capture complex relationships between input tokens. However, it underperformed on the anxiety task, potentially due to overfitting or noise sensitivity. This reveals that while powerful, RoBERTa might require additional regularization or task-specific adaptation for short, informal text.

5.3 Fusion Strategy and Multimodality

The fusion approach used in all experiments was early fusion, where visual (CLIP) and textual embeddings were concatenated before classification. This strategy, while simple and efficient, may not be optimal for capturing nuanced inter-modal interactions. For example, a caption like "this is fine" paired with a chaotic image carries a meaning far more intense than what either modality conveys individually. Future models may benefit from attention-based cross-modal interaction mechanisms that allow the text and image to influence each other dynamically during representation learning.

5.4 Challenges in Figurative and Sarcastic Language

Meme content is rife with figurative language, sarcasm, exaggeration, and cultural references. These linguistic phenomena pose a significant challenge for even the best language models. Our placeholder figurative reasoning module was not used in training, but its integration with large generative models like GPT-4 or a trained sarcasm detection head may improve understanding of implied meaning. Current encoders tend to take text literally, leading to misclassifications in cases where the surface text contradicts the emotional intent.

5.5 Error Patterns and Label Confusions

Qualitative analysis of misclassified memes revealed several trends:

- Memes with minimal or unreadable text often led to incorrect predictions, highlighting OCR's importance and its limitations under poor contrast or font styles.

- In the depression dataset, models frequently confused closely related symptoms (e.g., Anhedonia vs. Fatigue), suggesting a need for better intra-class discriminative features.

- In the anxiety dataset, memes expressing general unease were often classified as Nervousness, even when more specific labels like Excessive Worry were appropriate. This reflects the limitations of training data balance and semantic granularity.

5.6 Dataset Limitations and Future Improvements

The datasets used in this study are well-curated but relatively limited in size and diversity. Some categories are underrepresented, contributing to skewed training and lower macro-F1 scores. Data augmentation techniques, such as paraphrasing captions or generating synthetic memes, may help in future experiments. Incorporating external commonsense knowledge sources (e.g., ConceptNet, ATOMIC) could also enhance the model's ability to infer implicit emotional states and symbolic associations.

5.7 Takeaways

Overall, our analysis suggests:

1. Simpler models like Sentence-BERT are effective for short-form classification when label structure is simple.

2. More expressive models like RoBERTa are better suited for complex, multi-label tasks but require careful tuning.

3. Fusion strategies and multimodal reasoning architectures must evolve to better capture nuanced inter-modal meaning.

4. Domain-specific pretraining (e.g., MentalBERT) provides a solid foundation, but adaptation to meme-specific context is crucial.

6 Conclusion and Future Work

Our proposed multimodal system demonstrates promising results in classifying depression and anxiety symptoms in memes. By combining OCR-extracted text with image features and integrating commonsense reasoning, we outperform basic baselines. In the future, we plan to:

- Incorporate emotion recognition in visuals to capture tone.
- Experiment with cross-attention mechanisms for better fusion.
- Extend the system to predict severity levels (e.g., mild, moderate, severe).

This work contributes to building empathetic AI systems capable of detecting mental health signals in online content, potentially assisting in early intervention strategies.