Mental Health Meme Classification

Het Riteshkumar Shah
IIIT-Delhi
het22213@iiitd.ac.in

Rahul Bharti
IIIT-Delhi
rahul22387@iiitd.ac.in

Symantak Paliwal
IIIT-Delhi
syamantak22529@iiitd.ac.in

Abstract

Building on the M3H framework and AxiOM dataset, we aim to enhance mental health meme classification by improving context-aware representations. While M3H achieves notable gains, our proposed refinements focus on better capturing figurative language and commonsense knowledge. These advancements aim to improve model interpretability and generalizability in real-world applications.

1 Introduction

Our work builds on prior research in three key areas: social media memes, mental health expressions online, and multimodal approaches to mental health analysis. Memes serve as a widely-used medium for communication, with studies exploring their categorization, spread, and impact. Social media's effect on mental health remains complex, with findings highlighting both risks and benefits, along with efforts to predict mental health conditions using online data. Multimodal methods integrate textual, visual, and behavioral cues for improved mental health analysis. Using the RESTORE dataset, which annotates depressive symptoms based on PHQ-9, we enhance state-of-the-art models by jointly analyzing visual and textual commonsense reasoning to capture implicit connections often missed by traditional multimodal techniques.

2 Related works

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 infor-

mation 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 Methodology

n this work, we propose replacing cosine similarity with an attention-based retrieval mechanism to better capture the relevance of OCR-extracted text in meme classification. Memes frequently contain text that is stylized, distorted, or embedded in complex visual backgrounds, which makes it difficult for traditional similarity measures, such as cosine similarity, to accurately assess the relationship between the text and the image.

To address this, we incorporate an attention mechanism that allows the model to focus on the most relevant parts of the extracted text based on the content of the image. In our approach, the visual features extracted from the meme are used to guide attention over the text features. This enables the model to assign higher importance to text segments that are more informative or contextually related to the image, while reducing the influence of irrelevant or noisy text.

This method improves the way visual and textual information are combined, leading to more accurate interpretations of the meme's message, including subtle elements such as sarcasm or implied meaning. By learning which parts of the text are most relevant in different visual contexts, the model is better equipped to classify memes that rely on complex interactions between text and image.

4 Dataset

Depression Subcategories : Each meme can be associated with one or more subcategories of depression, requiring a multi-label classification approach.

Anxiety : Each meme is assigned exactly one label from a predefined set of anxiety-related categories, making it a single-label classification task.

5 Experimental Setup

5.1 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:** This refers to recognizing the specific psychological state being expressed in the meme. Understanding the underlying emotional or cognitive state can provide critical insights into the meme's message.

5.2 Knowledge Fusion

To overcome the limitation of static models being unable to utilize dynamic artifacts, we fuse external knowledge into the M3H framework through a Retrieval-Augmented Generation (RAG) module. This integration enables more contextual and up-to-date comprehension of memes, ensuring that M3H remains informed by the most relevant and recent information.

We utilize training images I along with their corresponding figurative reasoning r and OCR text o, such that

$$I_t \in \{(o_1, r_1), \ldots, (o_n, r_n)\}.$$

These pairs are embedded to form the knowledge base as follows:

$$E = \Pi([o_1, o_2, \ldots, o_n])_{n\times d} \oplus \Pi([r_1, r_2, \ldots, r_n])_{n\times d}, \quad (1)$$

where Π denotes a sentence-transformer that computes linear embeddings for each tuple $\langle o_x, r_x \rangle$. The RAG database is constructed by concatenating embeddings of both OCR text and figurative reasoning, yielding the final RAG embedding matrix $E \in R^{n\times 2d}$.

For any new meme input, we retrieve the top-n most relevant meme indices γ from the database. This is done by computing an embedding for the new meme as: $\mathrm{e}_k = \Pi(o_k) \oplus \Pi(r_k) \in R^{1\times 2d}$, $\gamma = \arg\max_i \Phi(e_k, E_i)$, where the embedding e_k is compared against each entry E_i in the database using cosine similarity Φ, and the top-n most similar instances are retrieved. These retrieved instances provide additional contextual support for downstream tasks such as meme classification or reasoning.

5.3 Classifier

Once the relevant examples are retrieved, we integrate all the information to construct a coherent input for our classifier. This input is carefully designed to assist the classifier in accurately identifying mental health-related concerns depicted in memes. By leveraging knowledge fusion, our method enhances the classifier's understanding of meme content, with a focus on the text modality to identify psychological symptoms ,we inculcate the following attributes in our final input:

- **OCR-Text (o):** The text extracted from the meme, representing the literal or explicit content. This may include captions, dialogue, or any embedded textual elements and serves as an objective component in the classification process.

- **Figurative Reasoning (r):** A commonsense-driven interpretation of the meme's figurative language. It encapsulates elements such as humor, irony, and metaphors, which are frequently used to express complex emotional or mental states.

- **Relevant Knowledge Attributes (κ):** Contextually relevant information retrieved from the RAG module, which serves as external knowledge. This supplementary content improves the classifier's ability to interpret nuanced or abstract meme content by grounding it in similar historical examples.

6 Experiments and Results

Table 1: Performance Comparison on RESTORE and AxiOM Datasets

Metric	RESTORE Dataset	AxiOM Dataset
Baseline Macro F1	65.82%	64.45%
Baseline Weighted F1	65.85%	64.95%
RoBERTa Macro F1	66.78%	68.76%
RoBERTa Weighted F1	67.38%	69.29%

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.