Towards Identifying Fine-Grained Depression Symptoms from Memes

Shweta Yadav[*], **Cornelia Caragea**[*], **Chenye Zhao**[*], **Naincy Kumari**[†], **Marvin Solberg**[⋆], and **Tanmay Sharma**[‡]

[*]Department of Computer Science, University of Illinois Chicago, USA
[†]Central University of Rajasthan, India
[⋆]Wayne State University, USA
[‡]Indian Institute of Technology Gandhinagar, India
[*]{shwetay,cornelia,czhao43}@uic.edu, [†]knaincy818@gmail.com,
[⋆]marvin.solberg@wayne.edu, [‡]tanmay.sharma@iitgn.ac.in

Abstract

The past decade has observed significant attention toward developing computational methods for classifying social media data based on the presence or absence of mental health conditions. In the context of mental health, for clinicians to make an accurate diagnosis or provide personalized intervention, it is crucial to identify fine-grained mental health symptoms. To this end, we conduct a focused study on *depression disorder* and introduce a new task of identifying fine-grained depressive symptoms from memes. Toward this, we create a high-quality dataset (RESTORE) annotated with 8 fine-grained depression symptoms based on the clinically adopted PHQ-9 questionnaire. We benchmark RESTORE on 20 strong monomodal and multimodal methods. Additionally, we show how imposing orthogonal constraints on textual and visual feature representations in a multimodal setting can enforce the model to learn non-redundant and de-correlated features leading to a better prediction of fine-grained depression symptoms. Further, we conduct an extensive human analysis and elaborate on the limitations of existing multimodal models that often overlook the implicit connection between visual and textual elements of a meme.

1 Introduction

Mental health disorders have a profound impact on society. Almost 1 billion people worldwide suffer from mental health disorders, predominantly depression, anxiety, mood, and substance use disorders (WHO, 2022). Two of the most common mental health disorder, depression and anxiety account for the US $1 trillion in economic losses worldwide annually (Health, 2020). This cost is projected to rise to a staggering US $6 trillion by 2030 (Bloom et al., 2012). Apart from the economic burden, the social burden of mental health disorders is huge. Suicide is now the fourth leading cause of death among those aged 15 to 29 years old (Fleischmann et al., 2021). However, considering its preponderance and global burden, depression continues to be significantly undertreated in all practice settings, where fewer than one-third of adults with depression receive effective treatment. Denial of illness and stigma are the two most common obstacles to appropriate diagnosis and treatment of depression (Sirey et al., 2001).

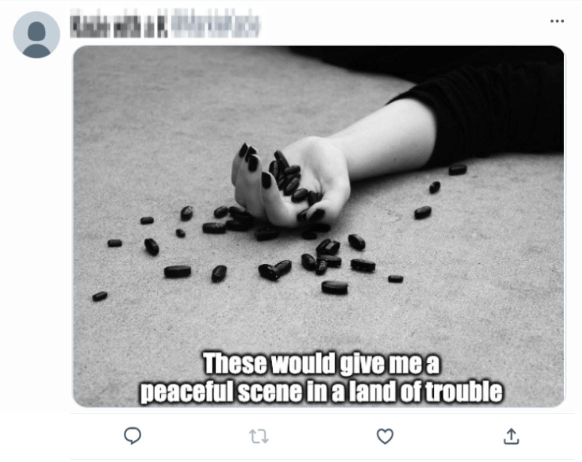


Figure 1: Example of a depressive meme. If we merely evaluate the textual content ("*These would give me a peaceful scene in a land of trouble*"), it is difficult to establish the author's true feelings. However, the meme-image can provide complementary information to help recognize the depression symptom (*self-harm*) correctly.

Recently, social media data has emerged as a powerful "lens" for tracking and detecting depression (De Choudhury et al., 2013; Yates et al., 2017). The vast majority of the existing works on depression have utilized the textual or multi-modal information available in the social media data to primarily classify the posts based on the perceived depressive behavior (depressive or non-depressive) (De Choudhury et al., 2014; Coppersmith et al., 2015; Gui et al., 2019). However, for healthcare professionals to intervene and provide effective treatment, it is crucial for them to understand the leading symptoms of that depressive behavior.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8890–8905
July 9-14, 2023 ©2023 Association for Computational Linguistics

Motivated by this, we aim to develop a practical decision support system that swift through social media posts and can provide healthcare professionals deeper insights into one's depressive behaviors by capturing the fine-grained depressive symptoms. In the past, there have been few attempts (Yadav et al., 2020; Yazdavar et al., 2017) to capture the depression symptoms; however, they are confined to only textual information. Recently, a new form of communication has emerged in social media: *'meme'*. A meme usually consists of an expressive image embedded with a short block of text. It is designed to convey a complex idea or emotional state of mind, which is far easier to understand than a textual description of thoughts. They acknowledge a shared experience between the creator and the viewer and therefore have become a fast way of communication on social media. Outside the mental health domain, there are numerous studies on meme processing and understanding for emotion detection (Sharma et al., 2020; Pramanick et al., 2021a), cyberbullying (Maity et al., 2022), and hateful meme detection (Zhou et al., 2021; Pramanick et al., 2021b).

However, to our knowledge, none of the existing studies have yet leveraged the visual information available in memes specifically to capture the fine-grained depression symptoms. There are two main reasons to consider the visual information: **(i)** according to a recent survey[1], images appear in more than 42% of tweet; and **(ii)** textual information alone cannot capture the overall semantic meaning. For example, in Figure 1, considering only the text, "*These would give me a peaceful scene in a land of trouble*", would not be sufficient to identify the depressive symptoms or even to distinguish whether the meme is depressive or not. However, it is evident from the image that the meme expresses suicidal thoughts/intent. Therefore, in order to obtain a holistic view of a post and accurately determine the depression symptoms, it is necessary to take into account the visual information available in social media posts.

To this end, we propose a new task – **Fine-Grained Depression Symptom Identification from Memes** and hypothesize that leveraging the multi-modal information available in the memes can more effectively help identify depression symptoms from social media posts. Towards this, we utilize clinically established 9-scale Patient Health Questionnaire (PHQ-9) (Kroenke and Spitzer, 2002) depression symptoms categories to classify the depressive memes that we collected from two popular social media forums – Reddit and Twitter. In particular, we make the following contributions:

(1) We create a high-quality dataset (RESTORE) consisting of $9,837$ depression memes, annotated with 8 fine-grained PHQ-9 symptom categories: *Feeling Down, Lack of Interest, Eating Disorder, Sleeping Disorder, Concentration Problem, Lack of Energy, Low Self Esteem, and Self Harm.*
(2) We perform extensive experiments with 20 state-of-the-art monomodal and multimodal approaches to benchmark our dataset and introduce orthogonality constraints in a multimodal setting to incorporate multiple perspectives present in the meme.
(3) We conduct a thorough human analysis and highlight the major findings and limitations of the monomodal and multimodal models. Our best-performing model obtains an F1-Score of only 65.01, demonstrating the challenge involved with meme processing for depression symptom identification task, and we believe that our dataset will promote further research in this direction.

2 Related Works

Based on the data modalities, we categorize the existing works on depression detection as follows:

Language As highlighted in Fine (2006) study, people's thoughts are frequently reflected in their language, and the linguistic cues, such as informal language usage, first-person referencing, and greater usage of negative emotion words, generally typify psychiatric disorders (Ramirez-Esparza et al., 2008; Resnik et al., 2015). Numerous research in computational linguistics has modeled the language usage in mental health-related discourse to predict mental health states (Tsugawa et al., 2015; Harman and Dredze, 2014) and infer risk to various mental disorders using social media data (Benton et al., 2017b; Coppersmith et al., 2016; Huang et al., 2015; Yadav et al., 2018, 2021). Most of the earlier works utilized a feature-driven approach (Resnik et al., 2015; Karmen et al., 2015) to detect depression. Recently with the availability of multiple benchmark datasets (Yates et al., 2017; Coppersmith et al., 2015), existing methods are designed using neural models (Orabi et al., 2018; Zhang et al., 2020). While most of these existing work studies depression at a coarser level, there

[1] shorturl.at/nuFMQ

have been only a few efforts towards inferring depressive symptoms by analyzing the textual information in social media posts (Yadav et al., 2020).

Vision The visual information available in shared images offers valuable psychological cues for understanding a user's depression status. Previous studies (Girard et al., 2014; Scherer et al., 2013; Zhu et al., 2017) conducted in a clinical setting have established that certain non-verbal behaviors such as downward gaze angle, dull smiles, and shorter average lengths of a smile characterize depressive behaviors. Recently, with the popularity of photo and video-sharing social networking services such as Instagram have piqued the interest of researchers in investigating people's depressive behavior from their visual narratives. Reece and Danforth (2017); Manikonda and De Choudhury (2017) investigated the role of public Instagram profiles in identifying a depressive user.

Multimodal (Language+Vision+Speech) In recent years, there has been growing attention towards exploiting multimodal information such as speech, vision and text for depression detection (Valstar et al., 2013, 2014; Ringeval et al., 2018). Existing studies have devised several neural approaches to effectively combine the information from various modalities. For instance, Yin et al. (2019) utilized the hierarchical bidirectional LSTM network to extract and fuse the local video and audio features to predict the degree of depression. Gui et al. (2019) proposed a multi-agent reinforcement learning method for identifying depressive users. An et al. (2020) developed the topic-enriched multi-task learning framework that achieved state-of-the-art performance on multimodal depression detection tasks. In contrast to the above approaches, our study aims to find the fine-grained depression symptom from memes that have not yet been explored before.

3 Dataset Creation

In this section, we present a new benchmark dataset: RESTORE[2] for identifying fine-grained dep**R**essiv**E** **S**ymp**TO**ms f**R**om m**E**mes, that was created following a clinically-guided approach and includes contributions from medical informatics experts and psychologist at each phase.

3.1 Task Structure

Dataset Selection and Curation. We collect posts from two popular social media platforms: Twitter and Reddit. We chose these platforms as a data source because of their rising popularity among depressive users to publicly express their thoughts, moods, emotions, and feelings. This, coupled with the greater degree of anonymity, facilitates self-disclosure and allows users to be more truthful and open in sharing sensitive issues and personal life without fear of being embarrassed or judged. Thus these user-generated self-narratives provide low-cost, large-scale, non-intrusive data to understand depressive behavior patterns and outcomes outside the controlled clinical environment, both in real-time and longitudinally.

To ensure that we capture a broader spectrum of depressive behaviors, we use a domain-specific depression lexicon (Yazdavar et al., 2017). The lexicon contains depression-related terms from 8 symptom categories following the PHQ-9[3] questionnaire. We use the depression lexicon to collect tweets from Twitter public profiles that mention at least one of the words from the lexicon in their profile description. In a similar manner, we collect Reddit posts; however, we restrict ourselves to the following subreddits: "`Mental Health`", "`depression`", "`suicide watch`", "`depression memes`", "`eating disorder`", and "`sleeping disorder`".

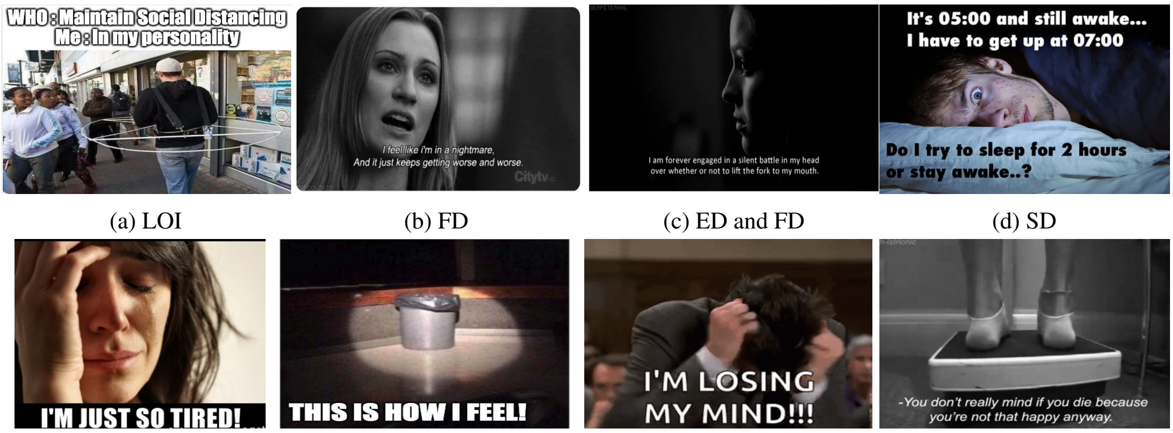
Objective. Given a meme (containing image and an embedded text) and an 8 fine-grained PHQ-9 depression symptom categories, the goal is to identify all depression symptoms that are expressed in the meme.

3.2 Task Construction

Filtering Strategy. Since the focus of this study is to map the content in memes to the corresponding PHQ-9 symptoms categories, we filtered out the posts that do not have a meme. Further, we applied a series of filtering steps to remove any irrelevant memes: **(i)** the meme should contain both image and embedded text (refers to the text which is embedded in the meme); **(ii)** the meme text must be written in English; **(iii)** the embedded text in

[2] The dataset is available at `https://forms.gle/qe5fYgesBZr1SiUd7`

[3] Though the PHQ-9 questionnaire includes 9 depression symptom categories, we excluded one symptom category, *"Hyper/lower activity"*, as this symptom can only be identified by the following behavior: "Moving or speaking so slowly that other people could have noticed?". Since this behavior can't be inferred by static social media data such as memes, we did not consider this symptom in our study.


WHO : Maintain Social Distancing
Me : In my personality
I feel like i'm in a nightmare,
And it just keeps getting worse and worse.
Citytv
I am forever engaged in a silent battle in my head
over whether or not to lift the fork to my mouth.
It's 05:00 and still awake...
I have to get up at 07:00
Do I try to sleep for 2 hours
or stay awake..?
I'M JUST SO TIRED!
THIS IS HOW I FEEL!
I'M LOSING
MY MIND!!!
-You don't really mind if you die because
you're not that happy anyway.

(a) LOI (b) FD (c) ED and FD (d) SD

(e) LOE (f) LSE (g) CP (h) SH and ED

Figure 2: Sample memes with associated PHQ-9 symptoms (LOI: Lack of Interest, FD: Feeling Down, ED: Eating Disorder, SD: Sleeping Disorder, LOE: Lack of Energy, LSE: Low Self Esteem, CP: Concentration Problem, SH: Self Harm)

the meme should be readable; **(iv)** the meme image should not be blurry and have a high resolution. Further, we filtered out those memes for which the OCR[4] could not obtain the text. Following these filtering criteria, we obtain $11,000$ posts.

Expert Annotation. We devised an annotation guideline based on the clinically adopted PHQ-9 depression questionnaire, which is a tool to assess the severity of depression. A team of 4 annotators (experts in psychology and medical informatics) independently annotated the collected memes. Each annotator was provided annotation guidelines and an interface to map the content in memes to the closest PHQ-9 symptom. Specifically, for a given meme, the annotators were instructed to label the depression symptom categories: *Lack of Interest, Feeling Down, Sleeping Disorder, Lack of Energy, Eating Disorder, Low Self-Esteem, Concentration Problem, and Self Harm*, that are the closest match to the meme based on the textual or visual information available in the meme. Note that symptoms can be one or multiple per meme, which renders the task as multi-label classification. If the meme does not contain any of these symptoms, annotators were instructed to label the meme in the *"Other"* class, which was not considered in our final dataset. For this task, the inter-annotator agreement (Krippendorff's alpha coefficient (Krippendorff, 2004)) is 81.55, which signifies a strong agreement amongst annotators. We provide examples for each symptom category corresponding to memes in Figure 2 and definition in **Appendix-A**.

3.3 Benchmark Dataset

Our final dataset includes $4,664$ depressive memes, and the distribution of PHQ-9 symptoms corresponding to these memes, as well as the train, test, and validation split, are shown in Table-1. Based on the obtained PHQ-9 class distribution, we can notice that a few PHQ-9 symptom categories are prominent in our human-annotated set, such as *'FD'*, *'ED'* and *'SH'*. In contrast, *'LOI'*, *'SD'*, *'LOE'*, and *'CP'* symptoms rarely appear in our human-annotated dataset.

To enrich and balance a few PHQ-9 symptom categories, we developed the training set with a portion of automatic curation. In our automatic curation of training samples, we followed two strategies to expand the human-annotated training set. In the first strategy, we conducted keyword-based search using *"eating disorder memes"*, *"feeling down memes"*, *"sleep disorder memes"*, *"lack of energy memes"*, *"low self esteem memes"*, *"concentration problem memes"*, *"self-harm"* on the Google Image and selected only top image search results. The second strategy considers selecting the top image results from Pinterest with the queries: *"insomnia memes"*, *"lack of interest memes"*, and *"sleep disorder memes"*. To remove noise, we maintained strict filtering on the resolution of the meme and on the readability of the meme's text. We also de-duplicate the memes if their sources are the same. Following this process, we obtained additional $5,173$ samples, which we used to enrich the training set. Also, it is to be noted that both the test and validation set only include manually annotated samples.

[4] https://cloud.google.com/vision/docs/ocr

Symptoms	LOI	FD	ED	SD	LOE	LSE	CP	SH	Total
Train	471	2085	1939	1562	122	855	595	1516	8814
Automatic	471	1070	454	1561	90	501	595	431	5173
Human	–	1015	1485	1	32	354	–	1085	3641
Test	97	294	98	99	95	128	100	106	662
Validation	45	195	49	45	54	85	42	61	361

Table 1: Data distribution in train, validation and test set for PHQ-9 symptoms. Both the test and validation set is human annotated.

4 RESTORE Dataset Analysis

Visual Analysis. We conducted a visual analysis of the memes to study how depression symptoms are related to color features. We performed color analysis by computing the pixel-level average w.r.t HSV (hue, saturation, and value), in which lower hue scores imply more redness and higher hue scores suggest more blue. Saturation describes an image's vibrancy. Value relates to the brightness of an image, with lower scores indicating a darker image. We observe that most of the memes, irrespective of symptom categories, are less colorful and have lower saturation values, suggesting negative emotions. These cases were prominent in *"low self esteem", "lack of interest"*, and *"self harm"*, where users often share memes that were less vivid, darker (higher grayscale), and have a high hue. In contrast, the memes related to *"eating disorder"* are brighter and more colorful, mainly because of the presence of food in the memes.

Qualitative Language Analysis. To understand the psycho-linguistics patterns associated with each PHQ-9 symptom category, we employed the LIWC (Tausczik and Pennebaker, 2010) to measure various linguistic factors such as *analytical reasoning, clout, originality, emotional tone, informal language markers, and pronouns*. Our analysis reveals that *"low self esteem"* has the lowest analytic reasoning among all the depression symptoms, depicting a more intuitive and personal language. Surprisingly, *"concentration problem"* has the highest analytic reasoning, suggesting formal and logical thinking patterns. The clout feature, which measures individual confidence and clarity in speaking or writing, was found to be highest in the *"feeling down"* and lowest in the *"eating disorder"* category. A similar trend was observed with the authentic feature, which is one way of presenting themselves to others in an original way. Further, we notice that individuals expressing *"self harm"* behavior, *"feeling down"*, and *"low self esteem"* symptoms use more first-person pronouns.

4.1 Benchmark Methods

We benchmark the RESTORE dataset on the following methods:

Monomodal (Language) Methods. We experiment with four pre-trained language models: BERT (Devlin et al., 2019), ROBERTA (Liu et al., 2019), XLNET (Yang et al., 2019) and MENTALBERT (Ji et al., 2021), fine-tuned on the RESTORE training set. For each model, we obtained the hidden state representations and utilized the feedforward network with the *sigmoid* activation function to predict the multi-label depression categories. Additionally, we also fine-tuned the BERT model by adding the LIWC features to integrate psycholinguistic information into BERT explicitly. We project the LIWC features using a feedforward network and concatenate these projected features with the BERT [CLS] token representation. The concatenated features are used to predict fine-grained depression symptoms. We call this network as the BERT+LIWC model.

Monomodal (Vision) Methods. To evaluate the effectiveness of visual information, we experiment with seven popular pre-trained vision models: DENSENET (Iandola et al., 2014), RESNET-152 (He et al., 2016), RESNEXT (Xie et al., 2017), CONVNEXT (Liu et al., 2022), REGNET (Schneider et al., 2017), EFFICIENTNET (Tan and Le, 2019), and VIT (Dosovitskiy et al., 2020). We fine-tuned these models on the RESTORE training set similar to the monomodal (language) models.

Multimodal Methods. We experiment with three state-of-the-art pre-trained multimodal models: VISUALBERT (Li et al., 2019), MMBT (Kiela et al., 2019), and CLIP (Radford et al., 2021), fine-tuned on the RESTORE training set. Additionally, we also experiment with the following models:

- **Late Fusion**: This model computes the mean prediction scores obtained from RESNET-152 and BERT model.
- **Early Fusion**: This approach concatenates features obtained from RESNET-152 and BERT, which are passed to a feed-forward network to make predictions.
- **BERT+HSV**: Here, we fine-tuned the BERT model by adding *mean*, *max*, and *min* values of HSV features of the image. Similar to BERT+LIWC, we concatenate HSV projected features with BERT [CLS] token representation to make predictions.

5 Proposed Approach

Existing multimodal approaches focus on generating text-image feature representations by detecting the objects in the image and learning an alignment between textual and visual tokens. However, a meme can convey multiple perspectives, and detecting the object alone may not be sufficient to generate a semantically-rich text-image representation. Therefore, to capture the image's multiple views that could be beneficial in effectively distinguishing the depression symptoms, we introduce orthogonal feature generation in a multimodal setting. We begin by first encoding the meme image $\mathcal{I}$ and its embedded text $\mathcal{T}$ with the pre-trained RESNET-152 model and BERT model, respectively. We selected these models because of their simplicity and comparable performance to other language-vision models. To capture multiple perspectives of the image, we perform the 2-dimensional adaptive average pooling (`adaptive-avg-pool`) of output size $S_0 \times S_1$ on RESNET-152 model output F that results in image representation $h_{\mathcal{I}} \in \mathcal{R}^{K \times S_0 \times S_1}$ of K feature map. With this approach, we obtained feature representations $h^1_{\mathcal{I}} \in \mathcal{R}^K$ and $h^2_{\mathcal{I}} \in \mathcal{R}^K$ by setting $S_0 = 2$ and $S_1 = 1$ (based on the validation performance).

Orthogonal Feature Generation: We introduce orthogonal feature generation, where the features are regularized with orthogonality constraints. With this constraint, we generate new features that capture another perspective, which are non-redundant and de-correlated with the existing features. The resulting orthogonal features help the model fully utilize its capacity, improving the feature expressiveness. Formally, given the textual feature $h_{\mathcal{T}}$ which corresponds to the BERT `[CLS]` token representation and image feature $h_{\mathcal{I}}$, we aim to generate the orthogonal feature $h_{\perp}$ to $h \in \{h^1_{\mathcal{I}}, h^2_{\mathcal{I}}, h_{\mathcal{T}}\}$ given another feature modality $\hat{h} \in \{h^1_{\mathcal{I}}, h^2_{\mathcal{I}}, h_{\mathcal{T}}\} - \{h\}$. Towards this, we first project the feature vector h into common vector space $\bar{h}$ thereafter, we compute the vector component C and orthogonal projection as follows:

$$C = \frac{\bar{h}^T \hat{h}}{\bar{h}^T \bar{h}} \bar{h} \quad \text{and} \quad h_{\perp} = \hat{h} - C \tag{1}$$

In this process, we obtained the orthogonal feature $h_{\perp}$ to $\bar{h}$ that also ensures (based on vector arithmetic) that it is non-redundant to $\hat{h}$.

Multimodal Fusion: In order to fuse both modalities, we devise a multimodal fusion strategy based on ***conditional adaptive gating***. Specifically, we first compute the bimodal scalars g^1 and g^2 with the gating mechanism (Rahman et al., 2020) by considering textual representation as one modality and one of the image features as another modality. These scalar values denote relevant information in the image feature conditioned on the textual feature. In the next step, we compute the multimodal representation considering both the image representation and the previously computed bimodal scalars with respect to the textual feature. Formally,

$$h_u = g^1 \mathbf{W}^1_f h^1_{\mathcal{I}} + g^2 \mathbf{W}^2_f h^2_{\mathcal{I}} \tag{2}$$

where $\mathbf{W}^1_f$ and $\mathbf{W}^2_f$ are weight matrices for both the image representation. With this strategy, we obtained the multimodal feature $f = h_{\mathcal{T}} + h_u$.

Depressive Symptoms Identification: Here, we first apply `LayerNorm` (Ba et al., 2016) operation on the multimodal feature f and orthogonal feature $h_{\perp}$. The resulting feature is concatenated with the textual feature $h_{\mathcal{T}}$ to form the final feature representation z. Finally, we apply the *sigmoid* operation on z to predict depression symptom categories.

6 Implementation Details

We utilized the pre-trained weights of BERT-base[5], RoBERTa-large[6], MentalBERT[7] and XLNet-base[8] from HuggingFace (Wolf et al., 2020). For the pre-trained vision models, we followed the `torchvision` API[9] and obtained the pre-trained weights of the vision models. In particularly, we use `resnet152`, `resnext101_32x8d`, `densenet161`, `efficientnet_b4`, `regnet_y_800mf`, `vit_l_32`, and `convnext_large` pre-trained weights to fine-tune on the PHQ-9 depression symptom identification task. We use the HuggingFace implementation[10] of VisualBERT to fine-tune the model on the PHQ-9 depression symptom

[5] https://huggingface.co/bert-base-uncased

[6] https://huggingface.co/roberta-large

[7] https://huggingface.co/mental/mental-bert-base-uncased

[8] https://huggingface.co/xlnet-base-cased

[9] https://pytorch.org/vision/stable/models.html

[10] https://github.com/huggingface/transformers/tree/main/examples/research_projects/visual_bert

identification task. For MMBT[11] and CLIP[12] also we follow the HuggingFace implementation and fine-tune the model on PHQ-9 depression symptom identification task. For the visual analysis of the RESTORE dataset, we use the `open-cv python` library[13]. We fine-tuned each model on the RESTORE training dataset for 10 epochs. The length of the maximum original text is set to 256 tokens. We normalized the images with pixel mean and standard deviation values before feeding them to the monomodal and multimodal networks. We evaluate the performance of each model on the RESTORE validation dataset and use the best (maximum micro F1-score) model to evaluate the performance on the RESTORE test dataset. To update the monomodal (vision) model parameters, we used AdamW (Loshchilov and Hutter, 2018) optimizer with the learning rate of $4e-5$. For the monomodal (language) and multimodal approaches, we used the AdamW optimization algorithm with a learning rate of $4e-5$. We set the batch size 64 to train all the benchmark models. We train the proposed network with batch size 16 and AdamW optimization algorithm (with the learning rate of $2e-5$) for 10 epochs. The dimension (K) of feature map obtained from RESNET-152 is 2048. For LIWC and HSV experiments, we set the size of the hidden unit as 20. We performed all the experiments on a single NVIDIA Tesla V100x GPU having 32GB memory. We observed the average runtime to train our framework is 11.55 minutes per epoch. The proposed model has $\sim$ 170 million parameters. All the libraries used in the experiment are licensed under the following:

- HuggingFace (3.5.0): Apache-2.0 License
- NLTK (3.6.3): Apache-2.0 License
- spacy (3.4.4): MIT License
- LIWC (22): Academic License
- open-cv (4.5.4): Apache-2.0 License
- PyTorch (1.10.1): modified BSD license

7 Results and Observations

Models	Precision	Recall	F1-score
BERT	$0.677_{0.01}$	$0.588_{0.01}$	$0.63_{0.01}$
XLNET	$0.647_{0.03}$	$0.577_{0.06}$	$0.609_{0.05}$
MLP+LIWC	$0.224_{0.04}$	$0.527_{0.07}$	$0.311_{0.02}$
BERT+LIWC	$0.684_{0.05}$	$0.557_{0.03}$	$0.613_{0.01}$
MENTALBERT	$0.695_{0.02}$	$0.579_{0.02}$	$0.631_{0.01}$
ROBERTA	$0.675_{0.07}$	$0.558_{0.01}$	$0.611_{0.03}$
DENSENET-161	$0.385_{0.0}$	$0.448_{0.01}$	$0.414_{0.01}$
RESNET-152	$0.368_{0.03}$	$0.425_{0.04}$	$0.394_{0.03}$
RESNEXT-101	$0.375_{0.01}$	$0.436_{0.01}$	$0.403_{0.01}$
CONVNEXT	$0.35_{0.08}$	$0.462_{0.01}$	$0.395_{0.04}$
MLP+HSV	$0.205_{0.02}$	$0.416_{0.05}$	$0.274_{0.02}$
REGNET	$0.377_{0.01}$	$0.427_{0.0}$	$0.4_{0.01}$
EFFICIENTNET	$0.315_{0.05}$	$0.449_{0.01}$	$0.369_{0.03}$
VIT	$0.36_{0.02}$	$0.404_{0.03}$	$0.38_{0.02}$
LATE FUSION	$0.637_{0.08}$	$0.58_{0.07}$	$0.601_{0.01}$
CONCAT BERT	$0.654_{0.05}$	$0.594_{0.04}$	$0.62_{0.01}$
BERT+HSV	$0.688_{0.05}$	$0.565_{0.04}$	$0.618_{0.01}$
VISUALBERT	$0.68_{0.03}$	$0.569_{0.03}$	$0.627_{0.01}$
MMBT	$0.66_{0.03}$	$0.58_{0.03}$	$0.616_{0.01}$
CLIP	$0.567_{0.13}$	$0.534_{0.08}$	$0.537_{0.01}$
PROPOSED	$0.693_{0.01}$	$0.607_{0.01}$	$0.651_{0.01}$

Table 2: Performance of the monomodal-language (first block), monomodal-vision (second block), multimodal (third-block) and proposed method on RESTORE test dataset. The MLP refers to multi-layer perceptron network. The reported results are the mean value of three runs with different random seed values. The subscript denotes the corresponding standard deviation.

Main Results Table 2 provides the summary of the results of monomodal and multimodal approaches. The obtained results (first two blocks of the table) show that pre-trained language models are better at capturing depression symptoms than the pre-trained vision models. We hypothesize that the existing vision models are pre-trained on generic IMAGENET (Deng et al., 2009) classes. Thus, these models lack the deeper semantic understanding of images that are required to effectively encode the memes' visual information in order to distinguish the depression symptoms categories precisely. While our finding reveals that the monomodal (language) model performs better than the monomodal (vision) model, we found that multimodal models having sophisticated fusion mechanisms such as VISUALBERT, and MMBT obtain significant improvement over the BERT on multiple symptom categories. This signifies that visual content is helpful in accurately classifying depression symptoms if used with a better mechanism to fuse the visual information with language features. Further, we observed that among all the competitive methods, our approach obtained the best performance in terms of the F1-score (*cf.* Table 3). For two classes (*SH* and *LOI*), MENTALBERT outperformed all the other models. We speculate that this

[11] `https://github.com/huggingface/transformers/tree/main/examples/research_projects/mm-imdb`

[12] `https://huggingface.co/openai/clip-vit-base-patch32`

[13] `https://docs.opencv.org/4.x/d6/d00/tutorial_py_root.html`

MODELS	LOI	FD	ED	SD	LOE	LSE	CP	SH	AVG
BERT (Devlin et al., 2019)	$0.369_{0.0}$	$0.732_{0.01}$	$0.812_{0.03}$	$0.742_{0.02}$	$0.047_{0.04}$	$0.41_{0.07}$	$0.788_{0.01}$	$0.556_{0.01}$	$0.63_{0.01}$
XLNET (Yang et al., 2019)	$0.329_{0.04}$	$0.718_{0.03}$	$0.777_{0.07}$	$0.726_{0.04}$	$0.084_{0.06}$	$0.395_{0.11}$	$0.754_{0.04}$	$0.534_{0.07}$	$0.609_{0.05}$
MLP+LIWC	$0.159_{0.14}$	$0.558_{0.07}$	$0.161_{0.04}$	$0.151_{0.12}$	$0.121_{0.11}$	$0.265_{0.05}$	$0.253_{0.04}$	$0.249_{0.05}$	$0.311_{0.02}$
BERT+LIWC	$0.366_{0.05}$	$0.71_{0.01}$	$0.823_{0.03}$	$0.749_{0.04}$	$0.055_{0.1}$	$0.383_{0.04}$	$0.726_{0.06}$	$0.557_{0.05}$	$0.613_{0.01}$
MENTALBERT (Ji et al., 2021)	$0.405_{0.01}$	$0.722_{0.02}$	$0.821_{0.02}$	$0.739_{0.03}$	$0.117_{0.02}$	$0.405_{0.07}$	$0.759_{0.01}$	$0.603_{0.03}$	$0.631_{0.01}$
ROBERTA (Liu et al., 2019)	$0.348_{0.05}$	$0.71_{0.02}$	$0.811_{0.04}$	$0.785_{0.06}$	$0.112_{0.01}$	$0.344_{0.05}$	$0.732_{0.03}$	$0.535_{0.06}$	$0.611_{0.03}$
DENSENET-161 (Iandola et al., 2014)	$0.143_{0.05}$	$0.611_{0.01}$	$0.386_{0.01}$	$0.414_{0.03}$	$0.0_{0.0}$	$0.184_{0.12}$	$0.467_{0.03}$	$0.295_{0.02}$	$0.414_{0.01}$
RESNET-152 (He et al., 2016)	$0.163_{0.06}$	$0.57_{0.04}$	$0.398_{0.04}$	$0.425_{0.04}$	$0.0_{0.0}$	$0.155_{0.11}$	$0.43_{0.06}$	$0.327_{0.03}$	$0.394_{0.03}$
RESNEXT-101 (Xie et al., 2017)	$0.052_{0.04}$	$0.608_{0.03}$	$0.403_{0.01}$	$0.355_{0.05}$	$0.0_{0.0}$	$0.131_{0.04}$	$0.406_{0.06}$	$0.304_{0.02}$	$0.403_{0.01}$
CONVNEXT (Liu et al., 2022)	$0.129_{0.16}$	$0.615_{0.01}$	$0.35_{0.07}$	$0.467_{0.03}$	$0.0_{0.0}$	$0.089_{0.15}$	$0.306_{0.28}$	$0.305_{0.04}$	$0.395_{0.04}$
MLP+HSV	$0.208_{0.09}$	$0.462_{0.09}$	$0.179_{0.16}$	$0.181_{0.03}$	$0.179_{0.12}$	$0.113_{0.1}$	$0.215_{0.07}$	$0.14_{0.14}$	$0.274_{0.02}$
REGNET (Schneider et al., 2017)	$0.064_{0.04}$	$0.596_{0.01}$	$0.386_{0.02}$	$0.424_{0.03}$	$0.0_{0.0}$	$0.094_{0.06}$	$0.449_{0.04}$	$0.31_{0.02}$	$0.4_{0.01}$
EFFICIENTNET (Tan and Le, 2019)	$0.019_{0.03}$	$0.624_{0.0}$	$0.308_{0.07}$	$0.338_{0.07}$	$0.0_{0.0}$	$0.005_{0.01}$	$0.119_{0.21}$	$0.278_{0.02}$	$0.369_{0.03}$
VIT (Dosovitskiy et al., 2020)	$0.107_{0.14}$	$0.601_{0.02}$	$0.368_{0.06}$	$0.263_{0.06}$	$0.0_{0.0}$	$0.087_{0.15}$	$0.21_{0.21}$	$0.22_{0.13}$	$0.38_{0.02}$
LATE FUSION	$0.355_{0.01}$	$0.704_{0.01}$	$0.72_{0.06}$	$0.72_{0.0}$	$0.02_{0.04}$	$0.308_{0.16}$	$0.79_{0.02}$	$0.543_{0.02}$	$0.601_{0.01}$
CONCAT BERT	$0.367_{0.0}$	$0.727_{0.02}$	$0.837_{0.01}$	$0.718_{0.04}$	$0.027_{0.05}$	$0.401_{0.05}$	$0.761_{0.02}$	$0.554_{0.03}$	$0.62_{0.01}$
BERT+HSV	$0.356_{0.02}$	$0.712_{0.03}$	$0.819_{0.01}$	$0.72_{0.05}$	$0.082_{0.07}$	$0.392_{0.05}$	$0.756_{0.01}$	$0.588_{0.04}$	$0.618_{0.01}$
VISUALBERT (Li et al., 2019)	$0.373_{0.01}$	$0.729_{0.01}$	$0.811_{0.01}$	$0.747_{0.03}$	$0.086_{0.03}$	$0.401_{0.05}$	$0.775_{0.02}$	$0.539_{0.02}$	$0.627_{0.01}$
MMBT(Kiela et al., 2019)	$0.374_{0.04}$	$0.716_{0.01}$	$0.842_{0.02}$	$0.747_{0.05}$	$0.033_{0.06}$	$0.411_{0.06}$	$0.697_{0.08}$	$0.555_{0.01}$	$0.616_{0.01}$
CLIP (Radford et al., 2021)	$0.247_{0.21}$	$0.668_{0.02}$	$0.696_{0.12}$	$0.617_{0.09}$	$0.013_{0.02}$	$0.22_{0.18}$	$0.675_{0.11}$	$0.457_{0.08}$	$0.537_{0.01}$
PROPOSED	$0.381_{0.01}$	$0.739_{0.01}$	$0.824_{0.01}$	$0.769_{0.02}$	$0.08_{0.03}$	$0.447_{0.06}$	$0.79_{0.01}$	$0.589_{0.02}$	$0.651_{0.01}$

Table 3: Class-wise performance of monomodal (language and vision), multimodal and proposed model on RESTORE test dataset.

Method	**Precision**	**Recall**	**F1-score**
Proposed Method	$0.693_{0.01}$	$0.607_{0.01}$	$0.651_{0.01}$
(–) Multimodal Fusion	$0.671_{0.01}$	$0.594_{0.01}$	$0.639_{0.01}$
(–) Orthogonal Feature	$0.686_{0.01}$	$0.568_{0.01}$	$0.625_{0.01}$
$h_\perp$ to $h_\mathcal{T}$ given $h_\mathcal{I}^1$	$0.673_{0.01}$	$0.598_{0.01}$	$0.637_{0.01}$
$h_\perp$ to $h_\mathcal{T}$ given $h_\mathcal{I}^2$	$0.661_{0.01}$	$0.602_{0.01}$	$0.628_{0.01}$
$h_\perp$ to $h_\mathcal{I}^1$ given $h_\mathcal{T}$	$0.693_{0.01}$	$0.607_{0.01}$	$0.651_{0.01}$
$h_\perp$ to $h_\mathcal{I}^1$ given $h_\mathcal{I}^2$	$0.658_{0.01}$	$0.608_{0.01}$	$0.632_{0.01}$
$h_\perp$ to $h_\mathcal{I}^2$ given $h_\mathcal{T}$	$0.671_{0.01}$	$0.601_{0.01}$	$0.641_{0.01}$
$h_\perp$ to $h_\mathcal{I}^2$ given $h_\mathcal{I}^1$	$0.667_{0.01}$	$0.594_{0.01}$	$0.639_{0.01}$

Table 4: Ablation results for the proposed approach.

is because a major portion of the corpus used to pretrain the MENTALBERT was centered on suicide and stress. For the *LOE* class, basic MLP+HSV model performs best because memes of these categories have higher grayscale and lower brightness values, which were effectively captured by HSV features. Though some of these approaches perform well in a particular depression category, they could not translate their performance across all the categories. In contrast, our proposed model shows competitive performance across most categories, which signifies the superiority of our proposed approach.

Ablation Study. To analyze the role of each component of the proposed method, we performed an ablation study and reported the results in Table 4 (top). We observe a performance drop of 1.2 and 2.6 points in the F1-score by removing multimodal fusion and orthogonal components. The significant performance drop confirms the importance of each component in predicting the symptoms category. We also analyze the role (Table 4, bottom) of imposing an orthogonal constraint on visual and textual features and find that feature orthogonal to $h_\mathcal{I}^1$ given $h_\mathcal{T}$ performs better compared to others.

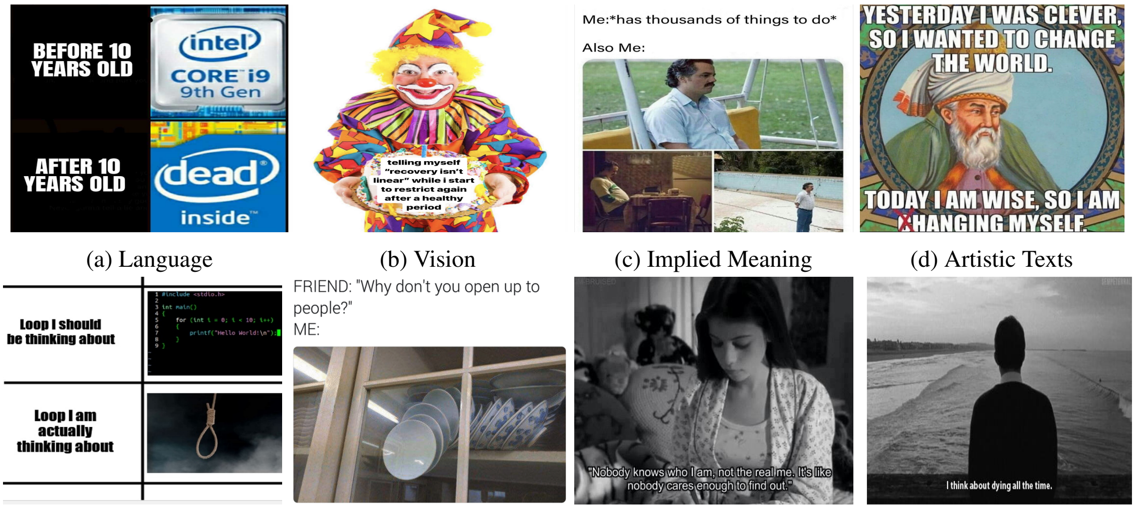
7.1 Analysis and Observations

We conducted an in-depth human analysis of models' predictions and came up with the following observations:

(a) Language. We noticed the memes that were correctly classified have clear depressive words. For example, consider Fig 3 (a), here the LM correctly predicted it as *'self-harm'* because of the presence of word *'dead'* in the text. This type of case was relatively higher for the classes, *'eating disorder'* and *'sleeping disorder'*.

(b) Vision. The vision models were also able to make correct predictions when a certain object in the meme correlated with the particular symptom class. For example, in Fig 3 (b) due to the presence of the *'cake'*, most of the models correctly predicted it as *'eating disorder'*.

(c) Implied Meaning. We observed that most of the models fail to infer an implicit sense of the memes. Fig 3 (c) shows an example of this error category made by all the models. Here, to correctly infer the depressive symptom, *'lack of interest'*, it is crucial to consider both the text and image which share complementary information. However, the multimodal models fail to judiciously fuse this complementary information leading to misclassification. The majority of the vision models predicted it as *'eating disorder'*, since the person is sitting on the dining chair and the models relates dining with


BEFORE 10 YEARS OLD
intel CORE i9 9th Gen
AFTER 10 YEARS OLD
dead inside
telling myself "recovery isn't linear" while i start to restrict again after a healthy period
Me:*has thousands of things to do*
Also Me:
YESTERDAY I WAS CLEVER, SO I WANTED TO CHANGE THE WORLD.
TODAY I AM WISE, SO I AM ~~C~~HANGING MYSELF.
Loop I should be thinking about
Loop I am actually thinking about
FRIEND: "Why don't you open up to people?"
ME:
"Nobody knows who I am, not the real me. It's like nobody cares enough to find out."
I think about dying all the time.

```
#include <stdio.h>

int main()
{
    for (int i = 0; i < 10; i++)
    {
        printf("Hello World!\n");
    }
}
```

(a) Language (b) Vision (c) Implied Meaning (d) Artistic Texts

(e) Figurative Speech (f) Figurative Speech (g) Generic Images (h) Generic Images

Figure 3: Human Analysis on the prediction obtained from monomodal and multimodal approaches.

eating.

(d) Figurative Speech. The usage of figurative speech is highly predominant in memes, mainly to compete with other memes and gain the attention and engagement of their followers. Our analysis reveals that both unimodal and multimodal models were not capable of dealing with figurative memes. For example, in Fig 3 (e), the word *'loop'* is used in the metaphoric sense, and neither the vision nor the LM understand the sense of the word *'loop'* or relate the 'rope' with the *'self-harm'*.

(e) Artistic Texts. Another way of making the meme more appealing to others is by using a variety of styling options on the texts. This brings a unique challenge for the OCR system to correctly extract all the text. For example, in Fig 3 (d), the OCR extracted the word *'changing'* instead of *'hanging'* leading to misclassification.

(f) Generic Images. We observed that few images which share the same aesthetic features do provide any symptom-specific visual cues. For example, in Fig 3 (g) and (h), if we just consider the image, we can only infer that person is feeling sad. It is in these cases the linguistic features are crucial in identifying the correct depression symptom class.

8 Conclusion

This work presents the first study towards identifying fine-grained depression symptoms from memes. We created a high-quality dataset – RESTORE, consisting of $9,837$ depressive memes annotated with PHQ-9 symptom categories and benchmark the dataset on 20 monomodal and multimodal models. Further, we introduce a novel method to incorporate various perspective in the meme that obtained best F1-Score over other approaches. Finally, our thorough human analysis of the model predictions indicates the model's limitation while dealing with memes, which will be considered in the future.

9 Limitations

This paper aims to make advancements toward automatically identifying fine-grained depressive symptoms from memes shared on social media. Although we used only those memes shared by users who self-declared themselves as depressive, we did not conduct any further clinical assessment to judge whether the user was depressive or not, nor we clinically evaluated their depression severity. Therefore, deploying this system without expert advice could compromise patient safety and lead to undesirable outcomes. We further acknowledge that determining the depression symptoms based on the visual and textual cues present in the meme can be subjective, and therefore the created gold-standard dataset may contain explicit and demographic biases. In this study, we focused on training the models using only the social media data, leaving their performance unchecked if tested on other medical data sources. Finally, our study is not indented to provide any diagnosis; instead, we envision the methods we provide being used as aids by healthcare professionals.

10 Ethical Consideration

Given that the created dataset is derived from social media and is focused on a sensitive mental health topic, we follow various ethical concerns regarding user privacy and confidentiality as inspired by (Benton et al., 2017a) to access and analyze the data. We adhere to the data usage privacy as provided by Twitter and Reddit to crawl the public profiles of their users. To ensure that we maintain the user's privacy, we anonymized the user profiles prior to the annotations, and we did not keep any meta-data information that would disclose the user. Further, we did not make any efforts to interact, deanonymize, or connect users on their other social media handles. The ethics review board approved the study under Human Subjects Research Exemption 4 because it is limited to publicly available social media posts. We believe that the created data would be highly beneficial to the community and to avoid any misuse (Hovy and Spruit, 2016), we will share the data with other researchers who will not deanonymize any of the users and will follow all the ethical considerations as established in this study.

References

Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal topic-enriched auxiliary learning for depression detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1078–1089.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017a. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017b. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162.

David E Bloom, Elizabeth Cafiero, Eva Jané-Llopis, Shafika Abrahams-Gessel, Lakshmi Reddy Bloom, Sana Fathima, Andrea B Feigl, Tom Gaziano, Ali Hamandi, Mona Mowafi, et al. 2012. The global economic burden of noncommunicable diseases. Technical report, Program on the Global Demography of Aging.

Mayo Clinic. 2022. Depression (major depressive disorder). Accessed: 2022-05-10.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 106–117.

Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM*, 13:1–10.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jonathan Fine. 2006. *Language in psychiatry: A handbook of clinical practice*. Equinox London.

Alexandra Fleischmann, Elise Paul, Devora Kestel, Bochen Cao, Jessica Ho, and Wahyu Retno Mahanani. 2021. Suicide worldwide in 2019.

Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. 2014. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 110–117.

GACCT Harman and Mark H Dredze. 2014. Measuring post traumatic stress disorder in twitter. *In ICWSM*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

The Lancet Global Health. 2020. Mental health matters. *The Lancet. Global Health*, 8(11):e1352.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562.

Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Christian Karmen, Robert C Hsiung, and Thomas Wetter. 2015. Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. *Computer methods and programs in biomedicine*, 120(1):27–36.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800.

Kurt Kroenke and Robert L Spitzer. 2002. The phq-9: a new depression diagnostic and severity measure.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes.

Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and understanding visual attributes of mental health disclosures in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 170–181. ACM.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

Shraman Pramanick, Md Shad Akhtar, and Tanmoy Chakraborty. 2021a. Exercise? i thought you said'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis. In *ICWSM*, pages 513–524.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Nairan Ramirez-Esparza, Cindy Chung, Ewa Kacewic, and James Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 2, pages 102–108.

Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):15.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 99–107.

Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. 2018. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pages 3–13.

Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE.

Nick Schneider, Florian Piewak, Christoph Stiller, and Uwe Franke. 2017. Regnet: Multimodal sensor registration using deep neural networks. In *2017 IEEE intelligent vehicles symposium (IV)*, pages 1803–1810. IEEE.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773.

Jo Anne Sirey, Martha L Bruce, George S Alexopoulos, Deborah A Perlick, Steven J Friedman, and Barnett S Meyers. 2001. Stigma as a barrier to recovery: Perceived stigma and patient-rated severity of illness as predictors of antidepressant drug adherence. *Psychiatric services*, 52(12):1615–1620.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3187–3196. ACM.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10.

WHO. 2022. World Mental Health Day: an opportunity to kick-start a massive scale-up in investment in mental health. Accessed: 2022-05-10.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709.

Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. 2018. Multi-task learning framework for mining crowd intelligence towards clinical treatment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 271–277.

Shweta Yadav, Usha Lokala, Raminta Daniulaityte, Krishnaprasad Thirunarayan, Francois Lamy, and Amit Sheth. 2021. "when they say weed causes depression, but it's your fav antidepressant": knowledge-aware attention framework for relationship extraction. *PloS one*, 16(3):e0248299.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198. ACM.

Shi Yin, Cong Liang, Heyan Ding, and Shangfei Wang. 2019. A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 65–71.

Yipeng Zhang, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, and Jiebo Luo. 2020. Monitoring depression trend on twitter during the covid-19 pandemic. *arXiv preprint arXiv:2007.00228*.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. 2017. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*, 9(4):578–584.

A PHQ-9 Depression Symptom Categories

Following are the depression symptom categories definition as provided by the Mayo Clinic (Clinic, 2022):

1. **Loss of Interest**: A decline in interest or pleasure in the majority or all normal activities, such as sexual activity, hobbies, or sports.
2. **Feeling Down**: Feelings of sadness, tearfulness, emptiness, or hopelessness.
3. **Sleeping Disorder**: Sleep disturbances, including insomnia, sleeping too much, or trouble falling or staying asleep.
4. **Lack of Energy**: Tiredness and lack of energy, so even small tasks take extra effort.
5. **Eating Disorder**: Reduced appetite and weight loss or increased cravings for food and weight gain.
6. **Low Self-Esteem**: Feelings of worthlessness or guilt, fixating on past failures or self-blame.
7. **Concentration Problem**: Trouble thinking, concentrating, making decisions, and remembering things.
8. **Self-Harm**: Frequent or recurrent thoughts of death, suicidal thoughts, suicide attempts, or suicide.

B RESTORE Dataset Analysis

B.1 PHQ-9 Symptom Co-occurrence.

Given that a single meme can have multiple depressive symptoms, we analyzed what symptoms occur together in a similar context through a co-occurrence heatmap, depicted in Figure 4. As can be observed, most of the samples had a single symptom. Only a few symptoms such as *“feeling down”* are more likely to occur with other symptoms, frequently with *“lack of self-esteem”*, *“self-harm”* and *“lack of energy”*. This is because these symptoms share common overlapping expressions with more generic *“feeling down”* symptoms. We also noticed for few cases where user expressing self-harm behaviors suffers from low self-esteem issues. This similar trend was also observed for eating disorder. Surprisingly, we observed a few uncommon co-occurrences, for instance, *“concentration problem"* and *“self harm"*.

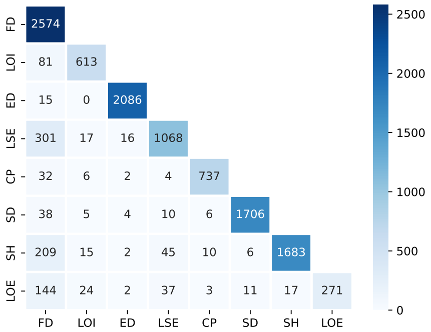


Figure 4: PHQ-9 Symptom Co-occurrence.

We have also provided the distribution of the memes with faces detected using Face++ API in Fig. 5. The study reveals that memes with *eating disorder* category contain a maximum of 60% faces and *sleeping disorder* memes contains 28% faces minimum amongst all the depression symptom category.

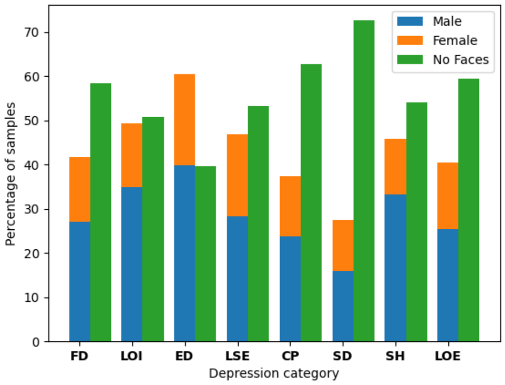


Figure 5: Distribution of the identified faces on respective depression symptom categories in the RESTORE dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
9

☑ A2. Did you discuss any potential risks of your work?
9

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3,4,5

☑ B1. Did you cite the creators of artifacts you used?
3,4,5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
6

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
6

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
10

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3,4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

4,5,6,7

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
6

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
6

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
7

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
6

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3, Appendix A

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Annotators are authors of the paper.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3,10

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
10

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Annotators are co-authors of this paper.