DynaMentA: Dynamic Prompt Engineering and Weighted Transformer Architecture for Mental Health Classification using Social Media Data

Akshi Kumar, *Senior Member, IEEE*, Aditi Sharma, *Member, IEEE*, Saurabh Raj Sangwan

***Abstract*—Mental health classification is inherently challenging, requiring models to capture complex emotional and linguistic patterns. Although large language models (LLMs) such as ChatGPT, Mental-Alpaca, and MentaLLaMA show promise, they are not trained on clinically grounded data and often overlook subtle psychological cues. Their predictions tend to overemphasize emotional intensity, while failing to capture contextually relevant indicators that are critical for accurate mental health assessment. This paper introduces DynaMentA (Dynamic Prompt Engineering and Weighted Transformer Architecture), a novel dual-layer transformer framework that integrates the strengths of BioGPT and DeBERTa to address these challenges. BioGPT captures fine-grained biomedical indicators, while DeBERTa provides context-aware disambiguation. The ensemble mechanism dynamically weights their outputs, guided by a simulated feedback loop that refines the predictions during training. Unlike previous studies that treat classification statically, DynaMentA incorporates dynamic prompt engineering to better align with evolving linguistic and emotional signals. Evaluated on three benchmark datasets, DepSeverity, SDCNL, and Dreaddit, DynaMentA achieves precision of 92.6%, 91.9% F1-score and 0.94 AUC-ROC, consistently outperforming the existing benchmark, including general-purpose LLMs and domain-specific mental health models. This scalable and interpretable framework establishes a state-of-the-art methodology for computational mental health analysis in high-stakes applications, such as suicide risk assessment and crisis intervention and early detection of severe depressive episodes.**

***Index Terms*—Mental Health Classification, Weighted Transformer Models, Social Media Data Analysis, Deep Learning for Social Systems, AI in Mental Health**

I. Introduction

Mental health disorders, including depression, anxiety, and stress, are critical global challenges, impacting millions of people annually and contributing significantly to the global burden of diseases [1]. According to the World Health Organization (WHO), depression alone is the leading cause of disability worldwide. Despite the prevalence of these conditions, early diagnosis and intervention remain inadequate due to the subjective nature of mental health assessments and the lack of scalable diagnostic solutions. The emergence of natural language processing (NLP) and artificial intelligence (AI) presents an unprecedented opportunity to bridge this gap. By analyzing textual data from online platforms such as social media posts, NLP models can uncover subtle linguistic cues indicative of mental health states [2]. Although AI-based mental health classification offers scalable solutions, it must be balanced with ethical considerations, including privacy, informed consent, and the risk of algorithmic bias in vulnerable populations. Moreover, existing approaches still struggle to capture the linguistic, cultural, and emotional diversity present in real-world mental health discourse.

Traditional methods, including rule-based systems and conventional machine learning approaches, face significant limitations. Rule-based systems often lack flexibility and do not account for linguistic variability, while traditional machine learning models rely heavily on extensive feature engineering and lack interpretability. Recent advances in large language models (LLMs), such as GPT- and BERT-based architectures, have shown great promise in mental health classification [3] [4]. However, these models are typically deployed in static configurations, relying on predefined prompts and isolated components, which hinder their ability to dynamically adapt to the evolving context of user input. This limitation affects their performance in nuanced tasks such as distinguishing between stress and depression or identifying suicidal ideation.

To address these challenges, this paper introduces **DynaMentA** (Dynamic Prompt Engineering and Weighted Transformer Architecture), a novel dual-layer transformer framework for mental health classification. The framework leverages two advanced transformer models: BioGPT [5], optimized to extract domain-specific cues from biomedical text, and DeBERTa [6], designed for context-aware classification using disentangled attention mechanisms and enhanced positional embeddings. Other transformer-based architectures, such as XLNet and T5, were considered but did not provide the same level of domain grounding or contextual precision required for mental health classification. DynaMentA's design uniquely fuses biomedical and contextual reasoning into a unified, interpretable framework tailored for emotionally complex text. Key innovations of **DynaMentA** include:

- **Dynamic Prompt Engineering:** A mechanism to generate adaptive prompts tailored to the linguistic and emotional context of the user, enhancing the extraction of relevant cues.
- **Weighted Ensemble Mechanism:** A task-specific ensemble that integrates the BioGPT and DeBERTa outputs,

Akshi Kumar is with the Department of Computing, Goldsmiths University of London, United Kingdom (e-mail: Akshi.Kumar@gold.ac.uk).

Aditi Sharma is with the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, India (e-mail: Aditi.Sharma@thapar.edu).

Saurabh Raj Sangwan with the Department of Computer Science and Engineering, Artificial Intelligence and Machine Learning, G L Bajaj Institute of Technology and Management, Greater Noida, India. (e-mail: saurabh.sangwan@glbitm.ac.in).

ensuring robust performance on binary and multiclass classification tasks.
- **Iterative Feedback Loop:** A simulated refinement mechanism that improves classification accuracy by iteratively adjusting the weighted ensemble outputs based on ground-truth labels.

This framework is rigorously evaluated on publicly available datasets, DepSeverity, SDCNL, and Dreaddit, demonstrating significant performance improvements with an accuracy of 92.6%, an F1-score of 91.9%, and an AUC-ROC of 0.94, outperforming state-of-the-art models.

The contributions of this paper are as follows:

- **Novel Framework Design:** A dual-layer transformer architecture combining BioGPT and DeBERTa for context-sensitive mental health classification.
- **Dynamic Prompt Engineering:** Implementation of user-specific adaptive prompts to improve linguistic cue extraction.
- **Weighted Ensemble Model:** Integration of a task-specific model that balances the contributions of the BioGPT and DeBERTa outputs to ensure optimal classification performance.
- **Comprehensive Evaluation:** Demonstration of significant performance gains in multiple datasets using rigorous evaluation metrics.

By addressing the critical limitations of existing methodologies, **DynaMentA** establishes a new benchmark in computational mental health analysis, offering a scalable, interpretable and ethically sound solution for real-world applications. The remainder of this paper is organized as follows. Section II discusses previous work on mental health classification using NLP and LLMs. Section III presents the formal problem statement and the learning objective. Section IV details the proposed DynaMentA framework, including its architectural components: BioGPT, DeBERTa, dynamic prompt engineering, a weighted ensemble mechanism, and a simulated feedback loop. Section V presents the experimental setup, including the datasets, baseline configurations, evaluation metrics, and the incorporation of specialized mental health models for comparative analysis. Section VI reports the results of these experiments, encompassing comparative performance analysis with standard and domain-specific models, error analysis, interpretability through attention heatmaps, and ablation studies evaluating the contributions of core components. Section VII addresses the ethical considerations of deploying AI systems in mental health contexts. Finally, Section VIII concludes the paper and outlines directions for future work.

II. Related Work

Over the last decade, social media has emerged as a valuable resource for understanding mental states and health trends [7] [8]. Studies have used linguistic patterns and social interactions on these platforms to identify risks such as anxiety, depression, and suicidal ideation [9] [10] [11]. Initially, prediction models used techniques such as SVMs using handcrafted features [12], but the rise of deep learning [13] has shifted the focus to pre-trained language models such as BERT, which have demonstrated effectiveness in mental health-related NLP tasks [14].

Transformer-based models have revolutionized NLP, with their highly parallel self-attention mechanisms [15]. Notable advances include BERT, GPT, and hybrid models like BART, each optimizing the pre-training and fine-tuning paradigm for diverse tasks [16] [17]. Larger models, such as GPT-3, have outperformed their predecessors, enabling capabilities such as few-shot learning [18]. Recently, ChatGPT and GPT-4 have gained significant attention for their human-like conversational abilities, while open-source models such as LLaMA have provided accessible alternatives for academic and industrial use [19].

In the health domain, LLMs have achieved remarkable results, particularly when fine-tuned in medical datasets [20]. MentalLLM [21] proposes a suite of instruction-tuned mental health models (for example, Mental-Alpaca, Mental-FLAN-T5), which outperform general-purpose LLMs like GPT-3.5 in task-specific evaluations using Dreaddit and SDCNL datasets. Some studies have used LLMs for sentiment analysis, emotion reasoning, and mental health classification tasks, but gaps in accuracy and performance persist [22]. Emerging models such as Mental-LLaMA [23] and initiatives such as the SMILE method [24] aim to address these challenges by expanding mental health datasets and improving LLM capabilities. Similarly, in addition, frameworks such as Psy-LLM [25] and Psycollm [26] integrate LLMs into mental health practice, offering real-time feedback to practitioners and advancing the role of AI in mental health support.

In this paper, we employ a weighted ensemble to combine BioGPT and DeBERTa with dynamic prompt engineering and iterative feedback to improve mental health classification.

III. Problem Statement

To formalize the problem, let $X = \{x_1, x_2, \ldots, x_n\}$ represent the set of user inputs, where x_i is an individual text sample (for example, a social media post). The goal is to assign to each input x_i a label $y_i \in Y$, where Y represents the set of mental health classes, such as $\{$Depression, Stress, Suicidal Ideation$\}$. The task can be framed as a supervised learning problem, where the objective is to learn a mapping function $f : X \to Y$ that minimizes the classification error. Mathematically, this can be expressed as:

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i) \tag{1}$$

where:

- $\mathcal{L}$ denotes the loss function (cross-entropy loss) for multiclass classification.
- f_θ is the parameterized model consisting of the BioGPT and DeBERTa components.
- θ represents the model parameters optimized during training.

The challenge lies in ensuring that f_θ generalizes well across diverse linguistic patterns and contexts. To achieve this, the proposed framework employs dynamic prompt engineering to enrich input representations x_i, with contextual cues C_i.

Here, C_i includes auxiliary features such as *temporal patterns* (e.g., posting frequency and timing relevant in **DepSeverity**), *syntactic structures* (for example, fragmented or telegraphic writing seen in **SDCNL**), and *lexical markers of emotional intensity or metaphorical language* (for example, "drowning," "trapped," frequently observed in **Dreaddit**). These cues are incorporated implicitly through dynamic prompt templates and guide both BioGPT and DeBERTa in adapting their embeddings to reflect task-relevant emotional and linguistic patterns.The final classification is performed using an ensemble method that combines predictions from BioGPT (f_{BioGPT}) and DeBERTa (f_{DeBERTa}):

$$\hat{y}_i = \alpha f_{\text{BioGPT}}(x_i, C_i) + \beta f_{\text{DeBERTa}}(x_i, C_i) \tag{2}$$

where α and β are task-specific weights satisfying $\alpha + \beta = 1$. In practice, α and β are derived by performing a grid search on combinations (for example, α from 0.1 to 0.9 in increments of 0.1, with $\beta = 1-\alpha$, selecting the pair that produces the highest F1-score in the validation set. Furthermore, the weighted sum approach was chosen over majority voting or stacking to preserve differentiability and allow task-specific emphasis during inference, facilitating adaptability across datasets with varying linguistic profiles. This innovative architecture sets a new benchmark in computational mental health analysis by ensuring that the system not only performs efficiently, but also adapts to the diverse and context-sensitive nature of mental health expressions.

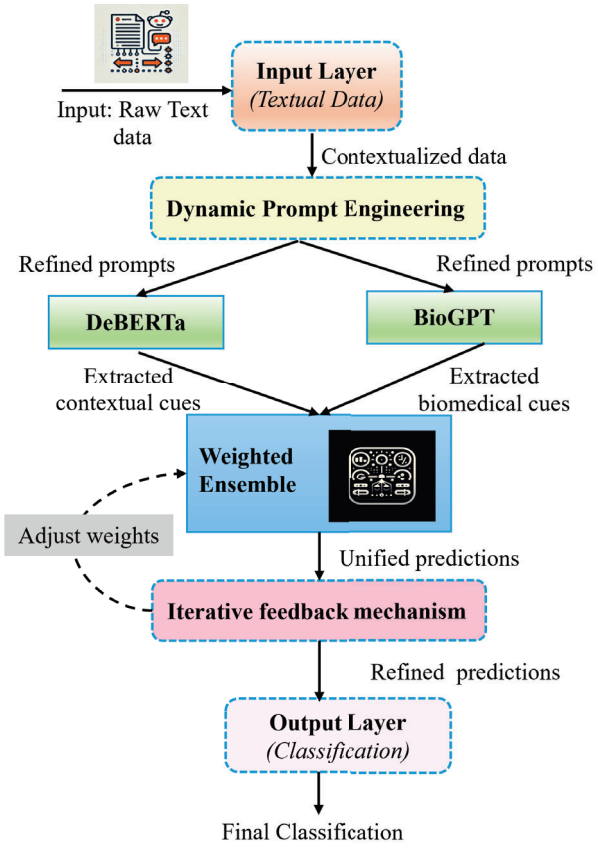

Fig. 1. The proposed DynaMentA Architecture

IV. Model Architecture

The proposed **DynaMentA**, Dynamic Prompt Engineering and Weighted Transformer Architecture, integrates two advanced transformer models, BioGPT and DeBERTa, to achieve high accuracy and robustness in mental health classification tasks. This architecture is designed to leverage the domain-specific capabilities of BioGPT for contextual cue extraction and the advanced classification abilities of DeBERTa for adaptive prediction. The two models work in tandem, their output being combined through a weighted ensemble mechanism to ensure optimal performance in binary and multiclass classification tasks.The high-level architecture of the proposed framework is illustrated in Figure 1.

A. BioGPT

The first component of the architecture, BioGPT, is a transformer-based language model pretrained on biomedical corpora, including PubMed abstracts, PMC-OA articles, and other publicly available biomedical datasets. The model utilizes self-attention mechanisms to extract domain-specific contextual cues from textual input, making it particularly effective in identifying linguistic patterns associated with mental health conditions. BioGPT processes an input sequence $X = \{x_1, x_2, \ldots, x_n\}$, where each x_i represents a tokenized segment of the input text. The output of BioGPT, $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, is a set of contextual cue vectors, each capturing semantic and syntactic information about the input text.

The processing involves a series of transformer layers, each consisting of multi-head self-attention and position-wise feedforward networks, ensuring that the model captures long-range dependencies and contextual nuances. For example, given an input such as *"I feel hopeless and can't see a way out"*, BioGPT extracts primary indicators like *suicidal ideation* and secondary indicators such as *hopelessness* and *entrapment*. These cues provide a structured representation of the user's mental state, which is crucial for downstream classification.

Dynamic prompts guide BioGPT in focusing on relevant linguistic patterns during contextual cue extraction. Unlike static prompts, which remain fixed irrespective of the input, dynamic prompts are generated adaptively based on the input context and historical interactions. The dynamic prompt mechanism is modeled as a function $P_t = f(C_t, H_t, \theta)$, where C_t represents the current context derived from the input, H_t are the historical interaction data and θ denote the learnable parameters of the prompt generation model. This mechanism ensures that BioGPT focuses on extracting the most relevant cues for the classification task at hand. For example, if a user input indicates work-related stress, the prompt dynamically adjusts to query stress-related details, such as *"Is this stress affecting your sleep or mood?"*. The ability to generate task-specific prompts significantly enhances the flexibility and contextual adaptability of the model.

B. DeBERTa (Decoding-enhanced BERT with disentangled attention)

The second component of the architecture, DeBERTa, serves as the classification layer of the framework. DeBERTa improves on traditional transformer architectures by incorporating disentangled attention mechanisms and enhanced relative positional embeddings. These features enable the model to capture fine-grained relationships between words and phrases, making it highly effective for understanding complex linguistic structures in mental health-related texts. DeBERTa processes the enriched input $\mathbf{x}_{\text{final}}$, which combines the original input sequence X with the contextual cues $\mathbf{V}$ extracted by BioGPT. The classification token $\mathbf{h}_{[CLS]}$ from the final transformer layer is passed through a task-specific feed-forward network to compute the probabilities of the class:

$$\hat{y} = \text{softmax}(\mathbf{W} \cdot \mathbf{h}_{[CLS]} + b), \tag{3}$$

where $\mathbf{W}$ and b are trainable weights and biases, respectively, and $\hat{y}$ represents the probability distribution over the output classes. For example, given an input such as *"I feel like nothing I do is ever good enough"*, the enriched input includes cues such as *low self-esteem* and *persistent doubt*. DeBERTa predicts a probability distribution that indicates the most likely mental health condition, such as *Depression* with a probability of 0.85.

C. Dynamic Prompt Engineering

To enhance adaptability across heterogeneous input distributions, we implement a dynamic prompt engineering module that conditions model behaviour on instance-specific linguistic and contextual variations. Rather than appending static task instructions, we introduce lightweight, input-aware prompt templates P_i that are programmatically constructed for each input x_i using derived auxiliary signals C_i.

Let x_i be a user-generated text sample, and C_i the set of contextual features extracted during preprocessing (e.g., part-of-speech patterns, negation cues, lexical affect, or time metadata). The prompt P_i is generated as:

$$P_i = \text{template}(C_i) \oplus x_i, \tag{4}$$

where $\text{template}(C_i)$ denotes a slot-filled prompt prefix constructed by mapping contextual features to semantic indicators (e.g., markers of temporal urgency, syntactic disfluency, or affect polarity), and $\oplus$ denotes prompt concatenation.

The prompts are injected into the tokenization layer for both BioGPT and DeBERTa and aligned with each model's tokenizer vocabulary to ensure embedding consistency. During training, templates are selected from a predefined pool and dynamically adjusted via rules or heuristics based on dataset-specific traits. No prompt tuning is performed at the embedding level; instead, the design leverages natural language scaffolds to condition attention flow and guide encoder behaviour implicitly.

This mechanism introduces inductive bias without introducing additional trainable parameters, supporting better generalization to edge-case inputs and reducing overfitting to lexical artifacts common in mental health corpora. Empirically, we observed improved performance on samples exhibiting metaphorical ambiguity or low-frequency constructs when dynamic prompts were used.

D. Weighted ensemble mechanism

The BioGPT and DeBERTa outputs are combined using a weighted ensemble mechanism, which balances the strengths of both models. The ensemble output is computed as:

$$y_{\text{ensemble}} = \alpha \cdot y_{\text{BioGPT}} + \beta \cdot y_{\text{DeBERTa}} \tag{5}$$

where α and β are task-specific weights satisfying $\alpha + \beta = 1$. These weights α and β are optimized during training using a grid search approach, ensuring that the contributions of BioGPT and DeBERTa align with the characteristics of the dataset and maximize the classification performance. These weights determine the ensemble output by appropriately balancing the predictions from both models, as demonstrated in the example. For example, if BioGPT predicts *Stress* with a probability of 0.7 and DeBERTa predicts *Stress* with a probability of 0.8, and the weights are $\alpha = 0.4$ and $\beta = 0.6$, the ensemble output is calculated as $y_{\text{ensemble}} = 0.4 \cdot 0.7 + 0.6 \cdot 0.8 = 0.76$, resulting in a final prediction of *Stress*. The fixed weighting strategy was chosen to balance the complementary strengths of BioGPT and DeBERTa while maintaining interpretability and reproducibility, which are critical in mental health applications. Although dynamic weighting strategies (for example, attention-based ensembling) may offer additional flexibility, we prioritized transparency and deterministic behaviour in this work.

E. Simulated Feedback Mechanism

The framework incorporates a simulated feedback mechanism to refine predictions and improve classification accuracy during training. In this study, the feedback is simulated using ground-truth labels from the datasets, as real-time user interactions are not available. If the model's prediction deviates from the ground truth, the feedback mechanism adjusts the ensemble weights to prioritize relevant cues extracted by BioGPT and DeBERTa. The iterative feedback loop updates weights only during training, refining model parameters based on misclassified instances to improve generalization to unseen test data. Therefore, the feedback mechanism is not deployed in real time but operates offline during training to adjust model behaviour based on known ground-truth labels. For example, if an input such as *"I feel exhausted and unable to focus"* is incorrectly classified as *Stress* instead of *Depression*, the feedback mechanism adjusts the ensemble weights to prioritize the depressive cues extracted by BioGPT and DeBERTa. While effective for controlled evaluations, this feedback mechanism currently does not incorporate user input in real time. In future iterations, the integration of human-in-the-loop feedback, such as clinician corrections or user-flagged misclassifications, could enhance the adaptability of the model and allow dynamic adjustment of ensemble weights or prompt construction based on real-world interactions.

Algorithm 1: DynaMentA: Dynamic Prompt Engineering and Weighted Transformer Architecture for Mental Health Classification

Require: Input text samples $X = \{x_1, x_2, \ldots, x_n\}$, model parameters $\theta_{\text{BioGPT}}, \theta_{\text{DeBERTa}}$
Ensure: Final predicted labels $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$
Preprocess X and split into training and test sets
for each training sample x_i **do**
  Extract contextual cues C_i
  Construct dynamic prompt $P_i \leftarrow \text{template}(C_i) \oplus x_i$
  Encode with BioGPT: $\mathbf{V}_i \leftarrow \text{BioGPT}(P_i; \theta_{\text{BioGPT}})$
  Compute prediction: $y^i_{\text{BioGPT}} \leftarrow \text{softmax}(\mathbf{W}_1 \cdot \mathbf{V}_i + \mathbf{b}_1)$
end for
for each training sample x_i **do**
  Concatenate input: $\mathbf{x}_{\text{final}} \leftarrow [x_i; \mathbf{V}_i]$
  Encode with DeBERTa: $\mathbf{h}_i \leftarrow \text{DeBERTa}(\mathbf{x}_{\text{final}}; \theta_{\text{DeBERTa}})$
  Compute prediction: $y^i_{\text{DeBERTa}} \leftarrow \text{softmax}(\mathbf{W}_2 \cdot \mathbf{h}_i + \mathbf{b}_2)$
end for
for each sample i **do**
  Combine predictions: $y^i_{\text{ensemble}} \leftarrow \alpha \cdot y^i_{\text{BioGPT}} + \beta \cdot y^i_{\text{DeBERTa}}$
  Final label: $\hat{y}_i \leftarrow \arg\max(y^i_{\text{ensemble}})$
end for
return Final predictions $\hat{Y}$

This dual-layer architecture ensures that the framework utilizes the domain-specific knowledge of BioGPT and the advanced classification capabilities of DeBERTa, achieving superior accuracy, robustness, and interpretability. The integration of dynamic prompts, weighted ensemble modelling, and simulated feedback further enhances the adaptability of the system, making it well suited for complex and context-sensitive tasks in mental health classification. The algorithm 1 summarizes the proposed dual-layer transformer framework for contextual mental health classification.

V. Experiments

To evaluate the performance of the proposed dual-layer transformer framework, we performed extensive experiments using publicly available datasets relevant to mental health classification. These datasets included annotated text samples containing user-generated content from social media platforms, forums, and clinical notes, categorized into binary and multiclass labels such as *Stress*, *Depression*, and *Anxiety*. Data were preprocessed by tokenization, lowercasing, and removal of stop words to ensure uniformity and reduce noise. The dataset was divided into training (80%), and test (20%) sets, maintaining class distribution across all subsets. The training dataset was further divided into a 70:10 ratio in the training and validation set. Computational experiments were conducted using NVIDIA Tesla V100 GPU with 32 GB of VRAM. The weighted ensemble mechanism and the simulated feedback loop were evaluated in ablation studies to measure their contribution to the overall framework.

A. Datasets

To ensure a robust and generalizable mental health classification, three publicly available datasets are utilized: **DepSeverity**, **SDCNL**, and **Dreaddit**. These datasets, sourced from Reddit, provide rich textual data that reflect real-world linguistic patterns and emotional tones, covering depression severity, suicidal ideation, and stress detection.

DepSeverity consists of 3,553 Reddit posts annotated into four levels of depression severity: minimal, mild, moderate, and severe. In accordance with the DSM-5 guidelines, it supports binary and multiclass classification tasks. For example, *"I feel like I can't do anything right anymore. Nothing makes me happy."* is labeled as *Moderate Depression*.

SDCNL (Suicide vs. Depression Classification Natural Language Dataset) contains 1,895 posts annotated as either *Suicidal Ideation* or *General Depression*. For instance, the post *"I don't think I can go on anymore. Life feels pointless."* is labeled as *Suicidal Ideation*.

Dreaddit comprises 1,191 posts focusing on stress detection in domains such as social and financial stress. The posts are labeled as stress or no stress, with additional metadata. For example, *"I'm behind on rent, and I don't know how I'll make ends meet."* is labeled as *Stress (Financial)*.

These datasets expose the framework to diverse linguistic patterns and emotional expressions, which enhances its robustness in handling mental health classification tasks. The preprocessing challenges included informal language, misspellings, and contextually ambiguous expressions, all of which could affect the accuracy of the classification. To mitigate these issues, we applied data normalization techniques such as slang replacement, spelling correction, stemming, and context-aware tokenization. These steps helped reduce noise and effectively structure the raw text for input into transformer-based models, improving overall input quality and model performance.

B. Baseline Models

To comprehensively evaluate the performance of the proposed DynaMentA framework, we compared it against three categories of baseline models: traditional classifiers, neural architectures, and state-of-the-art transformer-based language models.

Traditional Machine Learning Models. We implemented Support Vector Machines (SVM) with RBF kernel and Random Forests (RF) with 100 estimators as representative non-neural baselines. These models were trained on TF-IDF vector representations of the input text and serve to establish lower bounds on performance.

Neural Models. We included Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) to capture sequential and local features in the text, respectively. These models were initialized with pre-trained GloVe embeddings (300d) and trained using cross-entropy loss with Adam optimizer. Dropout and early stopping were used to prevent overfitting.

Transformer-Based Models. To benchmark against strong pretrained baselines, we fine-tuned BERT, RoBERTa, and XLNet using the Hugging Face Transformers library. All models were fine-tuned for 3–5 epochs using the same learning rate and batch size, with the final checkpoint selected based on the validation F1-score.

To ensure a fair comparison, all baselines were trained and evaluated on the same dataset splits and preprocessing pipeline.

C. Evaluation Metrics

Accuracy, precision, recall, F1-score, and AUC-ROC were selected as evaluation metrics to provide a balanced assessment of both the overall classification performance and the model's ability to handle class imbalances in mental health data. These metrics were chosen to reflect both general performance and sensitivity to misclassifications, which is critical in mental health prediction. Specifically, F1-score balances precision and recall, making it ideal for detecting underrepresented or overlapping classes. AUC-ROC provides a threshold-independent view of the model's discriminative power, especially important for high-stakes decisions involving stress, depression, or suicidal ideation. All metrics were computed on held-out test sets using standard scikit-learn implementations. Model performance was additionally validated with cross-validation to reduce evaluation bias.

D. Hyperparameter Settings

The training process used the AdamW optimizer with a cosine learning rate scheduler to dynamically adjust the learning rate throughout the training. A batch size of 16 was used and early stopping with a patience of 3 epochs was applied based on validation loss to prevent overfitting. Both BioGPT and DeBERTa were initialized with their respective pre-trained weights and fine-tuned on each mental health dataset. Hyperparameters, including learning rate, ensemble weights (α and β), dropout rate, and prompt construction parameters—were optimized using grid search. The learning rate was varied in the range 1×10^{-5} to 5×10^{-4}, and batch sizes of $\{8, 16, 32\}$ were explored. Prompt parameters included rule-based switches for template variation based on input metadata and emotional cues, rather than any learnable prompt embeddings. Ensemble weights were selected by maximizing F1-score on the validation set across pairs satisfying $\alpha + \beta = 1$. Hyperparameter combinations were evaluated on a held-out validation split consisting of 10% of the training data, and the configuration that generates the highest macro F1-score was selected for the final training. To ensure statistical robustness, all experiments were repeated five times with different random seeds, and mean and standard deviation of performance metrics were reported.

E. Experimental Inclusion of Specialized LLMs

To enhance the rigor of our experimental design, we further benchmarked DynaMentA against specialized large language models (LLMs) developed specifically for mental health prediction tasks. These include Mental-Alpaca and Mental-FLAN-T5 [21], as well as MentaLLaMA-chat-13B [23], which incorporate instruction tuning and psychological context to improve their relevance in mental health applications.

For fairness, the specialized models were evaluated on common benchmark datasets, Dreaddit, SDCNL, and Depression Reddit (DR) using consistent preprocessing, dataset splits, and evaluation metrics, including Balanced Accuracy and Weighted F1-score. Publicly released model checkpoints and tokenizers were used in inference mode to ensure reproducibility. Default settings were applied where specific hyperparameters were not available. The comparative performance outcomes of these models, alongside DynaMentA, are discussed in detail in the next section.

VI. Results and Discussion

The results demonstrate the significant advances achieved by the proposed DynaMentA in mental health classification. This section elaborates on performance metrics, dataset-wise results, ablation studies, error analysis, attention heatmaps, and training efficiency.

A. Comparative Performance Analysis

The proposed **DynaMentA** is compared with the baseline models, including SVM, LSTM, BERT, RoBERTa, BioBERT, and ChatGPT-4.0. Table I highlights the improvements in accuracy, F1-score, and AUC-ROC across all datasets.

TABLE I
Overall Performance Metrics

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC-ROC	MCC
SVM	79.6	78.9	78.1	78.5	0.81	0.56
LSTM	81.3	80.8	80.2	80.5	0.83	0.60
BERT	87.4	86.9	86.3	86.6	0.89	0.72
RoBERTa	88.1	87.6	87.1	87.3	0.90	0.74
BioBERT	86.2	85.8	85.0	85.4	0.88	0.70
ChatGPT	89.7	89.0	88.5	88.8	0.91	0.76
Proposed	**92.6**	**92.1**	**91.7**	**91.9**	**0.94**	**0.81**

DynaMentA consistently outperforms all baseline models across all metrics. While ChatGPT achieves high accuracy (89.7%) and F1-score (88.8%), DynaMentA surpasses it with a 3.2% higher accuracy and a 3.1% better F1-score. Compared to BioBERT, which specializes in biomedical text, DynaMentA demonstrates a 6.4% improvement in F1-score, demonstrating its ability to integrate domain-specific and contextual information effectively.

B. Dataset-Specific Comparison

The performance of **DynaMentA** was evaluated against several baseline models, including SVM, LSTM, BERT, RoBERTa, BioBERT, and ChatGPT-4.0, across three datasets: DepSeverity, SDCNL, and Dreaddit. Table II presents a detailed comparison of accuracy, F1-score, and AUC-ROC for each model on these datasets.

The results demonstrate that **DynaMentA** consistently achieves the highest performance in all datasets. For the SDCNL dataset, which focuses on the binary classification of suicidal ideation, DynaMentA achieves an F1-score of 92.5%, significantly outperforming ChatGPT by 3.2% and BioBERT by 6.7%. Similarly, in the DepSeverity dataset, which involves multi-class classification, DynaMentA achieves an F1-score of 90.7%, outperforming RoBERTa by 3.4% and BioBERT by 6.2%.

For the Dreaddit dataset, DynaMentA achieves an accuracy of 92.9% and an F1-score of 92.3%, highlighting its ability to handle nuanced sentiments and stress-related classifications better than ChatGPT, which achieves an F1-score of 90.6%.

TABLE II
DATASET-SPECIFIC PERFORMANCE COMPARISON ACROSS MODELS

Dataset	Model	Accuracy (%)	F1 (%)	AUC-ROC
DepSeverity	SVM	79.6	78.5	0.81
	LSTM	81.3	80.5	0.83
	BERT	87.4	86.6	0.89
	RoBERTa	88.1	87.3	0.90
	BioBERT	85.2	84.5	0.87
	ChatGPT	89.1	88.5	0.91
	DynaMentA	**91.3**	**90.7**	**0.93**
SDCNL	SVM	79.2	78.8	0.80
	LSTM	81.5	80.9	0.84
	BERT	87.7	86.8	0.89
	RoBERTa	88.7	87.9	0.91
	BioBERT	86.5	85.8	0.88
	ChatGPT	90.0	89.3	0.92
	DynaMentA	**93.1**	**92.5**	**0.95**
Dreaddit	SVM	79.5	78.7	0.81
	LSTM	81.8	80.7	0.84
	BERT	87.9	86.5	0.90
	RoBERTa	89.0	88.3	0.92
	BioBERT	87.0	86.4	0.89
	ChatGPT	91.2	90.6	0.93
	DynaMentA	**92.9**	**92.3**	**0.94**

C. Statistical Significance

To validate the observed improvements, a paired t-test was conducted between **DynaMentA** and the second-best performing model, ChatGPT-4.0, across all datasets. The results are summarized in Table III, where the p-values confirm the statistical significance of the improvement in DynaMentA performance ($p < 0.01$).

TABLE III
STATISTICAL SIGNIFICANCE OF DYNAMENTA'S IMPROVEMENTS

Dataset	Metric	Mean Difference	p-value
DepSeverity	F1-Score	2.2%	$p < 0.01$
SDCNL	F1-Score	3.2%	$p < 0.01$
Dreaddit	F1-Score	1.7%	$p < 0.01$

The improvements achieved by **DynaMentA** are not only statistically significant, but also practically relevant, as they consistently demonstrate the superior ability of the framework to handle both binary and multiclass mental health classification tasks in diverse datasets.

D. Comparison with Specialized Mental Health Models

Table IV summarizes the performance of DynaMentA compared to the specialized large language models on Dreaddit, SDCNL, and Depression Reddit (DR). The evaluation includes Mental-Alpaca, Mental-FLAN-T5 [21], and MentaLLaMA-chat-13B [23], which incorporate mental health–specific instruction tuning or domain adaptation, reflecting recent advances in task-specific language modelling.

DynaMentA consistently achieves higher scores in all datasets, outperforming specialized models by up to 15% in both accuracy and F1-score. These findings highlight the effectiveness of our dual-layer architecture in capturing both clinical semantics (via BioGPT) and contextual depth (via DeBERTa), making it better suited for cross-task generalization in mental health classification.

TABLE IV
PERFORMANCE COMPARISON OF DYNAMENTA AND SPECIALIZED LLMS ON OVERLAPPING DATASETS

Model	Dataset	Metric Type	Score (%)
Mental-Alpaca	Dreaddit	Balanced Accuracy	81.6
	SDCNL	Balanced Accuracy	72.4
Mental-FLAN-T5	Dreaddit	Balanced Accuracy	80.2
	SDCNL	Balanced Accuracy	67.7
MentaLLaMA-chat-13B	Dreaddit	Weighted F1	75.79
	DR	Weighted F1	85.68
DynaMentA	Dreaddit	Accuracy / F1-score	**92.9 / 92.3**
	SDCNL	Accuracy / F1-score	**93.1 / 92.5**
	DR	Accuracy / F1-score	**91.3 / 90.7**

E. Interpretability via Attention Maps

To enhance the interpretability of the proposed framework, attention heatmaps were generated for both BioGPT and DeBERTa. The heatmap represents the average attention weights across all heads for each layer. Figure 2 visualizes the BioGPT attention weights, highlighting words and phrases that contribute the most significantly to the prediction of the model. Similarly, Figure 3 illustrates the attention distribution for DeBERTa, providing information on how the model processes relational and contextual cues.

The attention heatmaps reveal the following patterns:

- **BioGPT** attends to biomedical and emotionally charged terms, such as *"hopeless," "exhausted,"* and *"trapped."* This focus enables the model to extract domain-specific cues highly relevant to mental health classification.
- **DeBERTa** emphasizes syntactic structure and relational context, attending to phrases such as *"can't see a way out"* or *"life feels pointless."* This enhances its capacity to refine predictions through contextual understanding.

This layered attention approach allows DynaMentA to capture both semantic and syntactic features, contributing to more explainable predictions. For example, in the sentence *"I feel hopeless and can't see a way out,"* BioGPT focuses on indicators such as *"hopeless"* and *"trapped"* while DeBERTa refines the classification by analyzing the broader relational context, particularly *"can't see a way out."*

Importantly, this example illustrates how attention heatmaps can enhance model transparency, allowing human observers, such as clinicians and system developers, to verify whether the model focus aligns with psychologically meaningful features. Interpretability is particularly critical in mental health applications, where understanding the rationale behind predictions can directly influence trust, adoption, and ethical deployment.

Table V provides examples of misclassified cases, helping to understand the challenges in nuanced classifications.

Errors often occur in overlapping linguistic patterns, such as between stress and depression, the model faces the ambiguity. For instance, "hopeless" is associated with both classes, leading to confusion. Incorporating additional features, such as

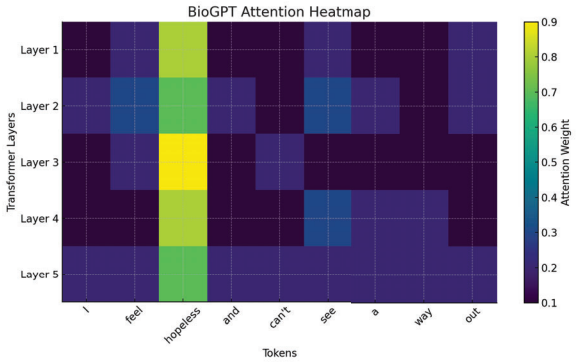

Fig. 2. BioGPT Attention Heatmap: Highlights critical biomedical terms contributing to predictions.

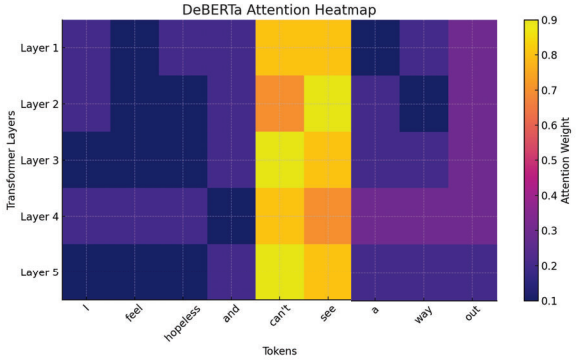

Fig. 3. DeBERTa Attention Heatmap: Emphasizes sentence structure and relational cues for refined classification.

TABLE V
ERROR ANALYSIS AND LIKELY CAUSES

Input Text	True Label	Predicted Label
"I feel tired and hopeless about work."	Depression	Stress
"Life feels pointless; I don't want to continue."	Suicidal Ideation	Depression
"I can't focus on tasks due to anxiety."	Stress	Anxiety

temporal data or multimodal inputs, could reduce such errors. To illustrate the limitations of attention-based interpretability, consider the following misclassified example:

"I just want a break from everything. My head feels like it's going to explode."

In this case, the ground truth label is **Stress**, but the model incorrectly classifies it as **Anxiety**. The attention heatmap in Figure 4 presents the BioGPT attention heatmap for a misclassified input. It shows that BioGPT focuses on emotionally intense tokens such as *"explode"* and *"everything"*, while underemphasizing the broader situational context that suggests external overload and burnout rather than internal fear or worry.

This misalignment suggests that the model can associate high-intensity emotional language with anxiety-like patterns, even when the underlying cause reflects acute situational stress. In addition, the absence of longitudinal or situational cues makes it difficult for the model to distinguish between transient stress and clinically significant anxiety. Such cases

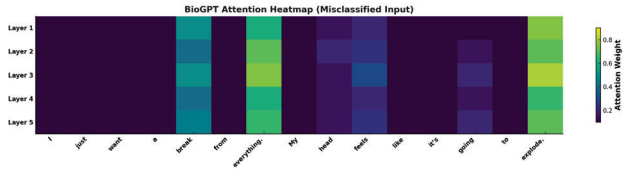

Fig. 4. BioGPT attention heatmap for a misclassified input.

underscore both the value and the limitations of attention heatmaps.

F. Performance by Dataset and Task Type

Table VI presents the framework performance in individual datasets for binary and multiclass classification tasks.

TABLE VI
PERFORMANCE BY DATASET AND TASK TYPE

Dataset	Task	Acc (%)	F1 (%)	AUC-ROC
DepSeverity	Multi-Class	91.3	90.7	0.93
SDCNL	Binary	93.1	92.5	0.95
Dreaddit	Binary	92.9	92.3	0.94

The framework performs consistently across datasets, with the highest accuracy of 93.1% on the SDCNL dataset, highlighting its effectiveness in distinguishing between suicidal ideation and depression. In the multiclass DepSeverity dataset, the framework achieves a high F1-score of 90.7%, demonstrating its ability to handle complex tasks involving multiple severity levels.

G. Ablation Study

To assess the importance of individual components in Dynamic Prompts and Weighted Ensemble Model Table VII presents results from the ablation study.

TABLE VII
ABLATION STUDY RESULTS

Configuration	Acc (%)	F1 (%)	AUC-ROC	MCC
Without Dynamic Prompts	89.2	88.7	0.91	0.76
Without Ensemble	90.4	89.8	0.92	0.78
Full Framework	**92.6**	**91.9**	**0.94**	**0.81**

Removing dynamic prompts reduced accuracy by 3.4%, indicating their critical role in extracting context-specific cues. The ensemble mechanism contributes an additional 2.2% improvement in accuracy, showcasing the synergy between BioGPT and DeBERTa.

H. Training Efficiency

The framework with 18 epochs converges within 12 epochs after reaching the threshold condition, demonstrating computational efficiency. Early stopping ensures that the model is not overfitting while maintaining high generalization performance. Figure 5 shows the training and validation loss curves, illustrating the convergence of the framework.

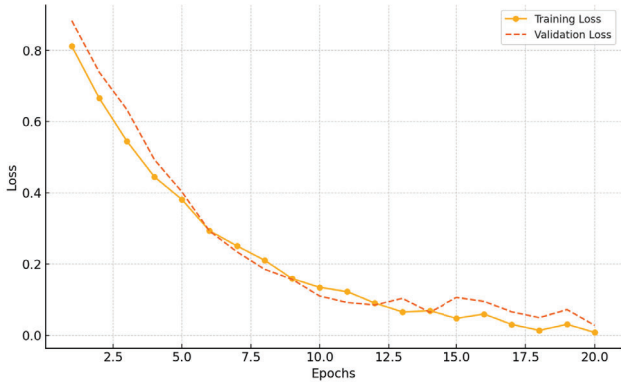

Fig. 5. Training and Validation Loss Curves.

VII. Ethical Considerations

The use of AI models for mental health classification raises several ethical concerns that must be carefully addressed:

- **Privacy and Data Security:** Mental health data is inherently sensitive. Ensuring robust mechanisms for data anonymization, encryption, and secure storage is paramount to prevent misuse or breaches.
- **Bias and Fairness:** Training data can contain inherent biases related to language, demographics, or cultural contexts. These biases could lead to skewed predictions, disproportionately affecting underrepresented groups. Rigorous testing and bias mitigation strategies are essential.
- **Informed Consent:** Collecting and using data for training purposes must be accompanied by explicit, informed consent of the individuals, ensuring transparency about how their data will be used.
- **Misuse and Overreliance:** AI models should not be used as standalone diagnostic tools. They must complement, not replace, professional mental health practitioners to avoid the risks of misdiagnosis or inappropriate interventions.
- **Transparency and Accountability:** Providing explanations for predictions is critical to building trust with users and practitioners. Models must include mechanisms for error reporting and accountability in case of inaccuracies or harm.
- **Ethical Deployment:** Deployment in real-world scenarios must align with ethical guidelines and regulatory standards, particularly in regions with stringent data protection laws such as GDPR or HIPAA.

By proactively addressing these ethical considerations, the framework can ensure its responsible development and deployment while prioritizing the well-being and trust of its users.

VIII. Conclusion

This paper proposes **DynaMentA**, a dual-layer transformer framework that combines BioGPT's biomedical expertise with DeBERTa's contextual modelling to address the complexities of mental health classification. Through dynamic prompt engineering and a weighted ensemble mechanism, DynaMentA adapts to diverse emotional and linguistic contexts, delivering robust predictions for both binary and multiclass tasks. An iterative feedback loop further enhances adaptability and reliability. Evaluations in the DepSeverity, SDCNL, and Dreaddit datasets demonstrate significant improvements in accuracy, recall, and F1-score, outperforming state-of-the-art baselines.

While **DynaMentA** has shown considerable advancements, it also has limitations. Its reliance on textual data restricts applicability to individuals who express mental health states through written language, excluding those who communicate verbally or non-verbally. Additionally, the use of Reddit-based datasets, although selected for their linguistic openness and relevance may limit generalizability between cultures and platforms. The simulated feedback mechanism, while effective during training, does not incorporate real-time user input, which is crucial for adaptive systems. Future work will address these challenges by extending the evaluation to other platforms (e.g., Twitter, Facebook, clinical data) and integrating multimodal inputs, such as audio, video and physiological signals for a more holistic understanding of mental health. Incorporating real-time human feedback and collaborating with mental health professionals will enhance clinical relevance, personalization, and trust. In particular, future development will explore human-in-the-loop mechanisms, where clinical input could refine predictions, guide prompt adaptation, and dynamically adjust ensemble weights. Furthermore, integrating user-specific language modelling in secure, consent-based environments (e.g., therapeutic chatbots) could enable more personalized and context-aware mental health support. These enhancements would bring DynaMentA closer to real-world deployment, ensuring greater inclusivity, adaptability, and societal impact.

References

[1] Y. Ibrahimov, T. Anwar, and T. Yuan, "Explainable ai for mental disorder detection via social media: A survey and outlook," *arXiv preprint arXiv:2406.05984*, 2024.

[2] M. Kerasiotis, L. Ilias, and D. Askounis, "Depression detection in social media posts using transformer-based models and auxiliary features," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 196, 2024.

[3] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, 2024.

[4] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin *et al.*, "Transformers and large language models in healthcare: A review," *Artificial Intelligence in Medicine*, p. 102900, 2024.

[5] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

[6] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[7] J. Kim, Z. A. Uddin, Y. Lee, F. Nasri, H. Gill, M. Subramanieapillai, R. Lee *et al.*, "A systematic review of the validity of screening depression through facebook, twitter, instagram, and snapchat," *Journal of Affective Disorders*, vol. 286, pp. 360–369, 2021.

[8] S. Dalal, S. Jain, and M. Dave, "Review of advancements in depression detection using social media data," *IEEE Transactions on Computational Social Systems*, 2024.

[9] J. H. Shen and F. Rudzicz, "Detecting anxiety through reddit," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, August 2017, pp. 58–65.

[10] F. T. Giuntini, M. T. Cazzolato, M. d. J. D. dos Reis *et al.*, "A review on recognizing depression in social networks: challenges and opportunities," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 11, pp. 4713–4729, 2020.

[11] M. Garg, "Towards mental health analysis in social media for low-resourced languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 3, pp. 1–22, 2024.
[12] S. Almutairi, M. Abohashrh, H. H. Razzaq, M. Zulqarnain, A. Namoun, and F. Khan, "A hybrid deep learning model for predicting depression symptoms from large-scale textual dataset," *IEEE Access*, 2024.
[13] M. A. Wani, M. A. ElAffendi, K. A. Shakil, A. S. Imran, and A. A. A. El-Latif, "Depression screening in humans with ai and deep learning techniques," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 2074–2089, 2022.
[14] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: a survey," *Pattern Recognition Letters*, vol. 167, pp. 204–211, 2023.
[15] A. Pandey and S. Kumar, "Mental health and stress prediction using nlp and transformer-based techniques," in *In 2024 IEEE Symposium on Wireless Technology Applications (ISWTA)*, July 2024, pp. 61–66.
[16] L. Ilias, S. Mouzakitis, and D. Askounis, "Calibration of transformer-based models for identifying stress and depression in social media," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1979–1990, 2023.
[17] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare," *arXiv preprint arXiv:2110.15621*, 2021.
[18] B. Lamichhane, "Evaluation of chatgpt for nlp-based mental health applications," March 2023. [Online]. Available: http://arxiv.org/abs/2303.15727
[19] M. Arcan, D.-P. Niland, and F. Delahunty, "An assessment on comprehending mental health through large language models," *arXiv preprint arXiv:2401.04592*, 2024.
[20] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large language models for mental health applications: Systematic review," *JMIR Mental Health*, vol. 11, no. 1, p. e57400, 2024.
[21] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-llm: Leveraging large language models for mental health prediction via online text data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–32, 2024.
[22] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis," *arXiv preprint arXiv:2304.03347*, 2023.
[23] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "Mentallama: Interpretable mental health analysis on social media with large language models," in *Proceedings of the ACM on Web Conference*, 2024, pp. 4489–4500.
[24] H. Qiu, H. He, S. Zhang, A. Li, and Z. Lan, "Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support," *arXiv preprint arXiv:2305.00450*, 2023.
[25] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang, "Psy-llm: Scaling up global mental health psychological services with ai-based large language models," *arXiv preprint arXiv:2307.11991*, 2023.
[26] J. Hu, T. Dong, L. Gang, H. Ma, P. Zou, X. Sun, D. Guo, and M. Wang, "Psycollm: Enhancing llm for psychological understanding and evaluation," *arXiv preprint arXiv:2407.05721*, 2024.

AUTHORS' BIOGRAPHIES

Dr. Akshi Kumar (Senior Member, IEEE) is the Director of Postgraduate Research and Chair of Research Ethics in the Department of Computing at Goldsmiths, University of London. With over a decade of experience in academia and research, her expertise spans Natural Language Processing (NLP), AI ethics, and explainable AI, with an expanding focus on health informatics and affective computing. Her work explores how emotionally intelligent and ethically grounded AI systems can support clinical decision-making, enhance mental health assessment, and advance digital health equity. In parallel, Dr. Kumar actively contributes to tackling misinformation, promoting digital integrity, and advancing AI literacy, particularly in the context of generative AI models and their societal implications. She has authored numerous high-impact journal articles and conference papers and is the author of Language Intelligence: Expanding Frontiers in NLP (IEEE-Wiley), a definitive guide for researchers and practitioners in the field. Dr. Kumar remains committed to the development of human-centered, transparent AI solutions for high-stakes domains such as healthcare, well-being, and public communication.

Dr. Aditi Sharma (Member, IEEE) is an Assistant Professor at Thapar Institute of Engineering and Technology, India. She has more than 5 years of experience in academia. Dr. Sharma received her Ph.D. from Delhi Technological University in the field of Affective Computing. Her research interests include Affective Computing, Machine Learning, Predictive Healthcare, and Text Summarization. She has multiple publications in high-impact SCI/SCIE journals. Dr. Sharma received the "Commendable Research Award for Excellence in Research" for her work at Delhi Technological University for the last three consecutive years. She holds professional memberships in IEEE, ACM, CSI, and IAENG.

Dr. Saurabh Raj Sangwan is an Assistant Professor in the Department of Computer Science and Engineering, Artificial Intelligence and Machine Learning, G L Bajaj Institute of Technology and Management, Greater Noida, India. He received his doctorate from Netaji Subhas University of Technology, New Delhi, in 2022. He completed his Bachelor's degree in Computer Science and Engineering from DCRUST, Murthal, India, and his M.Tech. Degree in Software Engineering from the Department of Computer Science & Engineering, Delhi Technological University, Delhi, India, in 2018. Dr. Sangwan's research has been published in high-impact journals and presented at reputed international conferences in the domains of data science and computational intelligence. He is a two-time recipient of the Commendable Research Award from NSUT, Delhi, recognizing his sustained scholarly contributions. His research interests include Health Data Mining, Cyber Informatics, and Natural Language Processing, with a particular emphasis on applying AI-driven methods to complex real-world problems in healthcare and and online safety.