LLM-Driven Psychological State Perception for Clinical Applications via Data Augmentation and Multidimensional Feature Fusion

Anonymous ACL submission

Abstract

Accurate perception of patients' psychological states is essential for personalized clinical decision-making and greatly depends on the effective classification of psychologically texts. However, existing approaches are often constrained by the scarcity of annotated data and the inherently implicit nature of psychological expressions. To address these challenges, we propose a unified framework that integrates large language model(LLM)–driven generative data augmentation with multidimensional feature fusion for effective psychological state perception. To alleviate data scarcity, a semantic-aware clustering strategy based on improved HDBSCAN is developed to obtain representative psychological semantic templates. These templates are incorporated as conditional prompts into the DeepSeek LLM to generate diverse and context-consistent psychological texts. To further enhance the diversity and quality of generated texts, emotional intensity and dialogue roles, and integrate original texts with multi-metric filtering, we propose a promotion evolution strategy by regulating emotional intensity and roles, combined with multi-metric filtering to ensure semantic consistency and reliability. To mitigate semantic ambiguity, we construct multi-grained representations that jointly model word-level clusters, sentence-level patterns, and cross-level semantic similarity. These features are encoded using TF-IDF and subsequently fused within a Naive Bayes classifier to achieve effective psychological state perception. Extensive experiments on multiple datasets demonstrate that the proposed method consistently outperforms state-of-the-art approaches, achieving a higher accuracy of 88.67%. The results highlight the effectiveness, robustness, and potential clinical applicability of the proposed framework.

1 Introduction

Patients undergoing medical treatment often express psychological distress through indirect and implicit linguistic patterns rather than explicit emotional statements. Such latent psychological factors can profoundly affect treatment adherence, risk of self-harm, and long-term prognosis. However, clinicians often struggle to accurately capture the psychological information due to time constraints in clinical consultations and the subtle nature of patient expressions. Therefore, the automatic perception of patients' psychological states from clinical texts, which supports clinicians in developing more personalized treatment plans, has emerged as one of the most clinically valuable applications of natural language processing in medicine.

The perception of psychological states based on text is essentially a text classification task. Text classification has evolved from traditional machine learning approaches to deep learning and, more recently, to pre-trained language models (PLMs). In the traditional machine learning stage, methods such as Naive Bayes and Support Vector Machines relied on manually engineered features, which constrained their generalization ability. In the deep learning stage, end-to-end learning is enabled by neural network architectures such as TextCNN (Lai et al., 2015) and LSTM (Liu and Guo, 2019), which automatically capture local semantic patterns and sequential dependencies. More recently, PLMs have introduced paradigm shifts in text classification, represented by models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ClinicalBERT (Alsentzer et al., 2019). By pre-training on large-scale corpora to learn general-purpose language representations and subsequently fine-tuning on downstream tasks, these models have achieved accuracy exceeding 95% on benchmark datasets and have become the dominant approach for contemporary text classification.

Despite the paradigm shift introduced by PLMs, their effectiveness deteriorates substantially when applied to clinical psychological texts. This limitation arises because psychological expressions are

rarely explicit; instead, they are highly contextualized, and their surface semantics often diverge from the underlying psychological intent. This challenge has motivated domain-specific advances, leading to models such as MentalBERT (Ji et al., 2022), PsychBERT (Vajre et al., 2021), and MentalLLaMA (Yang et al., 2024), which enhance the modeling of metaphors, somatic expressions, and emotional dynamics through continued pre-training on psychological counseling dialogues and depression forums. Building upon these efforts, the explicit incorporation of psycholinguistic resources, such as LIWC (Bahgat et al., 2022) and NRC-VAD (Mohammad, 2018), further improves sensitivity to cognitive process markers, self-referential language, and negation patterns. In parallel, architectural innovations such as the Hierarchical Attention Network (Yang et al., 2016) and cross-sentence relational reasoning mechanisms (Schuster et al., 2020) have been proposed to enable models to dynamically attend to salient symptom descriptions.

Psychological text classification for clinical decision-making has recently evolved from a standalone emotion recognition tool into a comprehensive clinical assistance system. To address the high heterogeneity and multidimensionality of disease diagnosis, researchers have proposed multimodal fusion approaches that integrate textual content, speech prosody, and physiological signals from wearable devices, significantly improving the detection of subtle emotional states (Al Hanai et al., 2018; Lin et al., 2020; Zou et al., 2022). To improve the standardization and consistency of medical reasoning, researchers have adopted knowledge graph–enhanced strategies that integrate diagnostic criteria from DSM-5-TR and ICD-11, as well as domain-specific knowledge bases such as MentalHealthKG, into model pre-training and inference pipelines via entity linking and relation embedding, reducing the risk of diagnostic hallucinations and improving the clinical reliability of model predictions (Gyrard et al., 2024).

However, patient expressions in medical scenarios exhibit greater specificity and concealment, while simultaneously facing stricter privacy protection requirements and data acquisition limitations. This renders psychological text classification in disease diagnosis and treatment still subject to three core challenges in practical applications. First, due to stringent ethical reviews and privacy safeguards, high-quality annotated data is extremely limited. Second, across different disease types and treatment stages, patients' psychological expressions show significant variations in linguistic patterns, emotional intensity, and thematic distributions. Consequently, traditional models lack sufficient generalization ability to adapt to such highly diverse contexts. Finally, when expressing negative emotions such as pain and fear, patients frequently employ euphemistic, implicit, or even deliberately concealing expressions, leading to a marked divergence between the underlying emotion and the literal meaning.

To address the aforementioned challenges, a patient psychological text classification method based on large language model-driven generative data augmentation and multi-dimensional feature fusion is proposed in this paper. The main contributions of this work are:

- A semantic template generation method to guide LLMs to generate new text is presented by incorporating psychological semantic perception and improved HDBSCAN clustering. The method involves promotion evolution for the LLM, rewriting the original text, and establishing a language quality evaluation method for text filtering to enhance data quality.

- We design a multi-grained psychological semantic fusion representation framework that jointly models word-level clustering distributions, sentence-level clustering patterns, word-sentence semantic similarity, and TF-IDF statistical features. This framework explicitly captures the implicit cross-level semantic relationships inherent in psychological texts, significantly improving classification and discrimination capabilities with small sample sizes.

2 Related work

For data augmentation, LLM-based generative methods (Chen and Shu, 2023; Chang et al., 2021; Sivarajkumar et al., 2024; Gao et al., 2025; Amugongo et al., 2025; Zhao et al., 2023) have been explored to alleviate data scarcity; however, existing methods still rely on coarse-grained conditional control, resulting in generated texts with lexical diversity only 60–70% of real data, with approximately 25–30% of samples containing logical contradictions. For unsupervised semantic analysis, clustering-based methods (Teferra et al., 2024; Shetty et al., 2020; Liu and Shi, 2022; Tavchioski

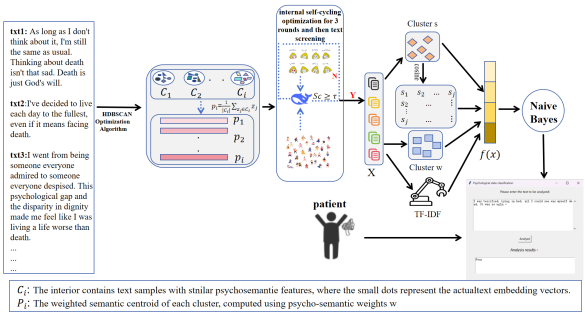

Figure 1: Overall algorithm framework of the proposed method.

et al., 2023; Lan et al., 2025; Rizwan et al., 2022; Le-Hinh et al., 2023; Guo et al., 2023) have been applied to mental health text partitioning and depression detection, but they primarily rely on shallow word embeddings and fail to provide fine-grained control signals for downstream conditional generation. For feature representation, fusion of TF-IDF with word embeddings (Onan et al., 2016), graph-based semantic relationship modeling (Nema et al., 2024), and multimodal integration have shown promise, yet most employ shallow concatenation strategies that are inadequate for capturing complex cross-level semantic interactions in psychological texts.

3 LLM-Driven Psychological State Perception

A patient psychological text classification strategy based on large language model data generation and multi-level feature fusion is proposed to effectively deal with data scarcity, implicit semantic expression, and contextual diversity. The overall framework includes three main parts: (1) A psychological semantic template for prompt words is obtained for LLMs-based data generation; (2) Evolution of text generation prompts for contextual diversity; (3) Design of a multi-level semantic feature fusion model to characterize psychological semantics.

3.1 Semantic Prompt Template Generation with Improved HDBSCAN

The quality of data generated by LLMs largely depends on how prompt templates precisely capture the semantics of psychological states. Manually designed templates tend to have limited expressiveness and are prone to introducing subjective bias. Although templates automatically generated by LLMs are comparatively objective and diverse, which often deviate from the true data distribution and introduce noise into subsequent recognition tasks.To address these issues, we leverage authentic psychological text data to develop an automatic semantic pattern identification method based on an improved clustering strategy.

HDBSCAN is selected as the base clustering algorithm due to its ability to operate without a predefined number of clusters and to capture density-connected regions of arbitrary shapes. We propose an improved HDBSCAN algorithm that incorporates semantic feature contributions across dimensions via a weighted cosine distance metric.

Let the true psychological text set of size N be $L = \{l_1, l_2, \ldots, l_N\}$. Each text is represented in D-dimensional space using BERT embeddings: $Z = \{z_1, z_2, \ldots, z_N\} \in \mathbb{R}^{N \times D}$. The contribution w_d of the d-th dimension's psycho-semantic features is defined as:

$$w_d = \sigma_d^2 \cdot s_d \tag{1}$$

where $\sigma_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_{i,d} - \bar{z}_d)^2$ is the variance of the semantic features in the d-th dimension, and s_d is the sentiment polarity measuring the correlation strength between the d-th dimension and gold-standard psychological sentiment semantics:

$$s_d = \frac{1}{|\mathcal{V}|}\sum_{w \in \mathcal{V}} \sigma\left(\frac{1}{N}\sum_{i=1}^{N} \cos(z_{i,d}, e_{\text{word}}) \cdot p_w\right) \tag{2}$$

Here, $\mathcal{V}$ is the mental lexicon obtained from the LIWC mental category lexicon; e_{word} is the BERT embedding of a lexicon word; $p_w \in \{-1, 1\}$ is the sentiment polarity annotation; and $\sigma(\cdot)$ is the sigmoid function.

Based on the contribution weights w_d, a weighted cosine distance is designed as:

$$d_{\text{wcos}}(z_i, z_j) = 1 - \frac{\sum_{d=1}^{D} w_d z_{i,d} z_{j,d}}{\sqrt{\sum_{d=1}^{D} w_d z_{i,d}^2}\sqrt{\sum_{d=1}^{D} w_d z_{j,d}^2}} \tag{3}$$

The psycho-semantic weighted mutual reachability distance (Campello et al., 2013) is defined as:

$$d_{\text{mreach}}(z_i, z_j) = \max\{d_{\text{core}}(z_i),\ d_{\text{core}}(z_j),\ d_{\text{wcos}}(z_i, z_j)\} \tag{4}$$

where $d_{\text{core}}(z_k)$ is the distance from sample z_k to its m_{pts} nearest neighbor:

$$d_{\text{core}}(z_k) = \min\{d \mid |\{z_l \neq z_k \mid d_{\text{wcos}}(z_k, z_l) \leq d\}| \geq m_{\text{pts}}\} \tag{5}$$

==========

Cluster 1 (Fear, p1):

Fear surged over me like a tide, and I couldn't help but imagine all kinds of possibilities for the future. Every thought left me feeling suffocated. I was afraid—afraid of losing the life I have now, afraid of becoming a burden to my family, afraid of facing the unforeseeable pain. Anxiety, like an invisible hand, clamped tight around my throat, robbing me of breath.

.....

==========

Cluster 2 (Positive and Optimistic, p2):

Life is all about mindset. It's your own life—face everything head-on, don't run away or complain, and meet it with positivity, Even though everything must end eventually, the journey can be uniquely your own

.....

==========

Cluster 3 (Fatalistic, p3):

Right now, my heart is like a windless lake, no ripples, no fear, not even a trace of sorrow. I know fate has pushed me onto a path with no turning back, one that's destined by heaven and beyond my control. Life is like a journey, and there's always a stop to reach—let it be.

Figure 2: Example of automatically generated psychological semantic templates.

Algorithm 1 Psychological Semantic Prompt Templates with Weighted Clustering

Require: Text set $L = \{l_i\}_{i=1}^{N}$, embedding dimension D, lexicon $\mathcal{V}$, m_{pts}, min_samples
Ensure: Clusters $\{C_i\}_{i=1}^{K}$, prototype vectors $\{p_i\}_{i=1}^{K}$
Encode $L \to Z \in \mathbb{R}^{N \times D}$ using BERT
for $d = 1$ to D **do**
  Compute σ_d^2, s_d, $w_d \leftarrow \sigma_d^2 \cdot s_d$
end for
for each pair (z_i, z_j) **do**
  Compute $d_{\text{wcos}}(z_i, z_j)$ via Eq. (3)
end for
for $k = 1$ to N **do**
  Compute $d_{\text{core}}(z_k)$ via Eq. (5)
end for
for each pair (z_i, z_j) **do**
  Compute $d_{\text{mreach}}(z_i, z_j)$ via Eq. (4)
end for
Apply HDBSCAN(D_{mreach}, m_{pts}, min_samples) $\to \{C_i\}_{i=1}^{K}$
for $i = 1$ to K **do**
  $p_i \leftarrow \frac{1}{|C_i|} \sum_{z_j \in C_i} z_j$
  Select sample in C_i with minimum d_{wcos} to p_i
end for
return $\{C_i\}_{i=1}^{K}$, $\{p_i\}_{i=1}^{K}$

For each valid cluster C_i, its semantic prompt template p_i is computed as the cluster centroid:

$$p_i = \frac{1}{|C_i|} \sum_{z_j \in C_i} z_j \tag{6}$$

The sample closest to p_i in the i-th cluster (by weighted cosine distance) is then selected as the semantic template. Example templates are shown in Figure 2. The pseudocode is presented in Algorithm 1.

3.2 LLM-based Psychological Text Generation with Evolved Prompts

We employ *DeepSeek* to generate psychological texts using carefully designed prompts and the semantic templates from Section 3.1. To address issues of repetitive and off-topic generation, we develop a prompt self-evolution method based on precise semantic control, as illustrated in Figure 3.

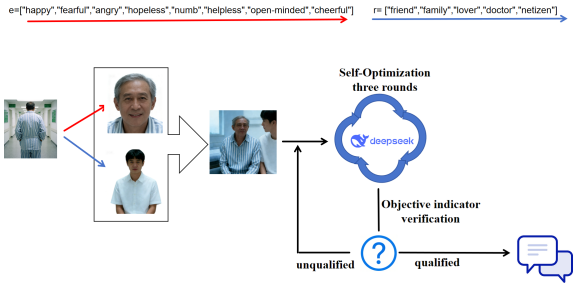

Figure 3: Text generation and filtering process.

The prompts incorporate patient-oriented emotions `e = {'happiness', 'fear', 'anger', 'despair', 'numbness', 'helplessness', 'open-mindedness', 'cheerfulness'}` and dialogue roles `r = {'friend', 'family', 'lover', 'doctor', 'online friend'}`. Each pair (e_i, r_j) generates diverse prompt variations to guide DeepSeek in generating new texts with consistent themes and diverse expressions. The full generation and filtering process is shown in Figure 3, with the evolutionary template in Figure 4 and an example regeneration trigger in Figure 5.

Three indicators are designed to evaluate the quality of generated psychological text:

PPL (Perplexity) A standard metric measuring the uncertainty of a model's prediction of a given text. A lower PPL value indicates more fluent and natural text. PPL is computed using a pre-trained GPT-2 model:

$$\text{PPL} = \exp\left(-\frac{1}{Q}\sum_{i=1}^{Q} \log p(q_i \mid q_{<i})\right) \tag{7}$$

TTR (Type-Token Ratio) Measures lexical diversity by calculating the ratio of unique words to total words. A higher TTR indicates richer vocabulary usage.

HDD (Hellinger Distance Divergence) Measures the statistical difference between the distribution of generated text and the reference mental text corpus (Hellinger, 1909). A lower HDD indicates better consistency with real mental discourse patterns.

A quality threshold vector $\boldsymbol{\tau} = [\tau_{\text{PPL}}, \tau_{\text{TTR}}, \tau_{\text{HDD}}]$ is defined. If any score fails to meet its threshold, the system automatically triggers a regeneration mechanism. Only high-quality texts meeting all thresholds are retained.

Context Type I: Caption

I want you to act as a patient. Your goal is to generate a new psychological text based on instructions and provided semantic templates, making it a bit difficult for those famous artificial intelligence systems to process. However, the rewritten text must be reasonable and must be understandable and responsive to humans.

Context Type II: Semantic Template Example

"Life is all about mindset. It's your own life—face everything head-on, don't run away or complain, meet it with positivity. Even though it all ends eventually, the journey can be uniquely yours."

Context TypeII: Constraint

emotion = ["happy", "fearful", "angry", "hopeless", "numb", "helpless", "open-minded", "cheerful"]
role = ["friend", "family member", "lover", "doctor", "online friend"]
###objective
...

Given Q&A

{
"Input": "Life is all about mindset; it's about facing everything head-on by yourself, not evading or complaining, but embracing it with positivity. Even though everything must end eventually, the journey can be uniquely your own."
"Emotion": Open-minded
"Dialogue Role": Lover
"Requirements": Facing {r} with mood {e}, imitate the input text to deeply express your psychological feelings upon learning of your cancer diagnosis. Do not describe physical reactions, and do not offer comfort or advice. Generate a piece of psychological text.
"Steps": First, generate a new piece of psychological text as required, then optimize it in terms of emotional depth, fluency, vocabulary richness and authenticity to get the final text. Then conduct an objective verification of the text; if it does not meet the predefined conditions, the system will trigger the above process to regenerate until a satisfactory output is obtained.
"Output": "I've come to see things clearly. I choose to face it bravely, no resentment, no regret for how life unfolds. Even though I can't walk the rest of the way with you, the time we've shared is already the most precious gift life has given us, and that's enough."## Rewritten Q&A
}

Figure 4: Evolutionary prompt template process.

```
DeepSeek Self-Assessment:
Overall score: 8.50/10
Emotional depth: 9.00/10
Smoothness:   8.50/10
Rich vocabulary: 8.00/10
Realism:   8.50/10
analyze: The metaphors are delicate, the emotional progression is natural, but the sentence
structures appear slightly ornate.
Local technical indicators:
PPL: 72.46 (Threshold: <45)
TTR: 0.852 (Threshold: >0.65)
HDD: 0.877 (Threshold: >0.75)

Quality check failed, triggering regeneration....
Local indicators not met
DeepSeek rating after regeneration: 8.50/10
After regeneration, PPL: 27.06

The text has been saved.: When the doctor said the word "normal," I thought I would feel relieved, but
my heart felt gripped tightly by an invisible hand, sinking into even deeper fear.
```

Figure 5: Example of regeneration mechanism triggered.

3.3 Multidimensional Feature Fusion

In psychological texts, the lexical level conveys fundamental emotional polarity and cognitive tendencies, whereas the sentence level integrates contextual information to form complete semantic expressions. To address the limitations of single-granularity representations, we construct multi-grained fused features.

Let the set of true and generated texts be $X = \{x_1, x_2, \ldots, x_n\}$ with n sentences total. We obtain t words via jieba segmentation and encode them with BERT, denoting word vectors as e_v and sentence vectors as e_s. K-Means clustering partitions all word and sentence vectors into H semantic clusters: $(C_v^h)_{h=1}^H$ and $(C_s^h)_{h=1}^H$ respectively, yielding semantic features f_v and f_s.

To capture the stable correspondence between local lexical themes and overall sentence patterns, the cosine similarity matrix between word and sentence cluster centers is computed:

$$M_{ij} = \cos\left(c_v^i,\, c_s^j\right) = \frac{(c_v^i)^\top c_s^j}{\|c_v^i\|\|c_s^j\|}, \quad i, j = 1, \ldots, H \tag{8}$$

Three statistics are extracted from M to quantify cross-granularity semantic consistency:

- **Maximum alignment** $M^{\max} = \max(M)$: indicates significant local-global semantic resonance.
- **Mean alignment** $M^{\text{mean}} = \frac{1}{|M|}\sum_{m \in M} m$: measures overall coordination between lexical topics and sentence patterns.
- **Alignment dispersion** $M^{\text{std}} = \sqrt{\frac{1}{|M|}\sum_{m \in M}(m - M^{\text{mean}})^2}$: quantifies the volatility of alignment degree.

To preserve fine-grained term information, conventional TF–IDF representation is incorporated as a supplementary feature. All multi-level features are concatenated into a unified representation and fed into a Naive Bayes classifier for psychological state inference.

4 Experiments

4.1 Experimental Background

To comprehensively evaluate the effectiveness of the proposed method, it is applied to three datasets. Among these, the first two are widely used public benchmark datasets for text classification, while the third is a self-constructed psychological text dataset of cancer patients.(1) **SimplifyWeibo_Mood (SWM)**: a ternary classification dataset comprising user comments from Weibo; (2) **MR**: a binary classification dataset of movie reviews;(3) $ZHIHU$ consists of psychological text data from cancer patients, collected from ZhiHu. All data in the $ZHIHU$ dataset were collected from publicly accessible online content and have been strictly anonymized before use. Any personally identifiable information, including usernames, IDs, locations, and other sensitive details, was removed or masked to ensure that the data cannot be traced back to any individual user. The statistical characteristics of the three datasets, i.e., number of documents, words, entities and classes, are summarized in Table 1. From Table 1, we can find that the number of texts is extremely different. The first two public datasets have 9100 and 10662 samples,

Table 1: Statistical characteristics of the three datasets.

	SWM	MR	ZHIHU
#Doc	9,100	10,662	380
#Word	118,300	18,764	8,318
#Entity	675,100	6,415	4,137
#Class	3	2	3

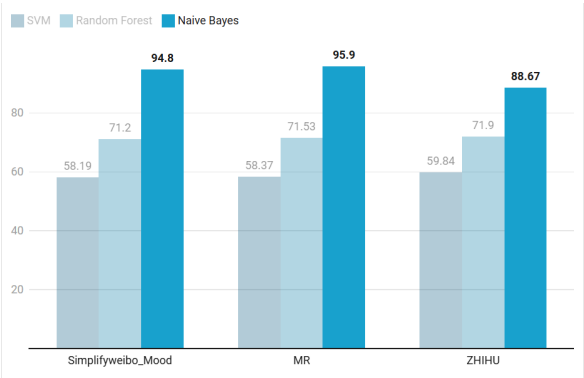

Figure 6: Classifier comparison on three datasets.

about 70 times larger than the third, indicating that psychology texts are really scarcity.

To simulate labeled data scarcity, 120 samples per category are randomly selected from each dataset as the training set, with the remaining samples used for testing. All experiments are conducted on an NVIDIA RTX 4060 GPU (CUDA 12.6), and evaluation metrics (accuracy and F1 score) are averaged over 10 repetitions.

4.2 Classifier Selection

Three classic classifiers—SVM, Random Forest, and Naive Bayes—are tested on the three original labeled datasets . As shown in Figure 6, Naive Bayes achieves the highest accuracy across all three datasets, improving accuracy by approximately 20% over SVM. Its simple probability-based assumptions are naturally suited to small sample sizes and high-dimensional sparse text data. Therefore, Naive Bayes is selected as the base classifier.

4.3 Baseline Comparison

The proposed text classification method based on generative data augmentation and multidimensional feature fusion using Naive Bayes classifier is compared with the following types of models to demonstrate its superiority. (1) Traditional Models: TF-IDF+SVM and LDA+SVM use TF-IDF features and LDA features to represent text, respectively, and then train SVM for classification. (II) Deep Learning Models: TextCNN, which uses pre-trained word vectors for initialization and extracts local n-gram features through a convolution neural network. (III) Pre-trained Language Models: Opinion-BERT and MentalLLaMA, where MentalLLaMA is a Chinese LLaMA model fine-tuned for the mental health domain, representing the current strongest domain-specific pre-trained model. To simulate the data scarcity, two sets, i.e., 120 per class training samples and 40 per class training samples, are tested here, and the results are listed in Table 2 and Table 3.

As shown in Table 2, the proposed GDA-MFF consistently outperforms all competing methods in terms of both accuracy and F1-score across the three datasets (Simplifyweibo_Mood, MR, and ZHIHU), whereas LDA+SVM yields the poorest results on all evaluation metrics. Specifically, GDA-MFF improves accuracy over the worst-performing baseline by 26.10%, 8.60%, and 27.47% on the three datasets, respectively; correspondingly, the F1-score gains are 28.00%, 9.20%, and 27.34%. This substantial performance gap can be primarily attributed to the inherent limitations of LDA, which relies solely on word co-occurrence statistics and is unable to capture deeper semantic relationships and contextual dependencies. Psychological texts are typically characterized by implicit semantics, strong contextual dependencies, and ambiguous topic boundaries, which pose significant challenges for LDA+SVM in accurately extracting latent psychological themes and consequently limit its feature representation capacity.

On the ZHIHU psychological dataset, GDA-MFF surpasses the second-best model, MentaLLaMA, by 6.37% in Accuracy and 5.54% in F1-score, further underscoring its superiority. Although MentaLLaMA demonstrates strong performance in psychological text classification, owing to its powerful semantic understanding enabled by large language models, it remains constrained by distribution shift and relatively homogeneous feature representations. These limitations become particularly pronounced in scenarios characterized by limited training data and subtle semantic expressions, both of which are prevalent in psychological text analysis. In contrast, GDA-MFF effectively addresses these challenges by integrating a LLM–driven data augmentation strategy with a multidimensional feature fusion mechanism. This design mitigates the inherent limitations of small-scale psychological datasets and facilitates more

Table 2: Results (%) of Accuracy and F1 Score on three datasets (120 samples per class).

Model	SWM		MR		ZHIHU	
	ACC	F1	ACC	F1	ACC	F1
TF-IDF+SVM	73.80	72.50	90.70	90.10	66.40	64.10
LDA+SVM	68.70	66.40	87.30	86.50	61.20	58.90
TextCNN	89.64	88.91	94.28	93.87	71.83	69.74
Opinion-BERT	71.30	70.70	95.10	94.70	76.80	78.90
MentaLLaMA	93.20	92.80	95.40	95.10	82.30	80.70
GDA-MFF	**94.80**	**94.40**	**95.90**	**95.70**	**88.67**	**86.24**

Table 3: Results (%) of Accuracy and F1 Score on three datasets (40 samples per class).

Model	SWM		MR		ZHIHU	
	ACC	F1	ACC	F1	ACC	F1
TF-IDF+SVM	53.85	45.18	51.47	49.23	57.50	48.13
LDA+SVM	41.03	36.10	54.20	48.26	50.00	33.16
TextCNN	46.15	44.36	58.98	59.10	27.50	19.68
Opinion-BERT	30.77	26.30	51.71	50.58	32.50	28.79
MentaLLaMA	61.15	59.66	69.87	65.41	80.91	78.14
GDA-MFF	**65.78**	**65.32**	**71.53**	**70.82**	**86.40**	**85.24**

accurate modeling of latent psychological states and subtle semantic variations. As a result, GDA-MF yields consistent improvements in classification accuracy, generalization, and domain adaptability, thereby offering a more robust and efficient solution for psychological text analysis.

With per-class training samples falling from 120 to 40, TextCNN and Opinion-BERT suffer the sharpest performance drops. Across three datasets, their accuracy declines average 41.04% and 41.65%, with F1-score falling by 43.13% and 44.88%. This can be primarily attributed to the strong reliance of such deep learning models on large-scale training data; under low-resource conditions, inadequate parameter estimation results in substantial performance degradation. In contrast, the proposed LLM-based template generation and data augmentation strategy proves effective. As a result, even under a substantial reduction in training samples, the performance of the proposed method on the ZHIHU dataset remains largely stable. However, for the other two datasets, the templates generated by the LLM fail to consistently align with the original data distribution in terms of semantic consistency and diversity. Consequently, the performance of the proposed method exhibits a noticeable decline on these datasets.

4.4 Ablation Study

To systematically evaluate the effectiveness of each module, ablation experiments are conducted on the following variants: (1) **RF**: baseline with no augmentation and no MFF; (2) **EV-D+MFF**: clustering-based template generation removed; (3) **CGST-D+MFF**: prompt evolution and filtering removed; (4) **MFF**: LLM-based text generation removed; (5) **GDA-MFF**: full proposed model. Results are shown in Table 4.

As shown in Table 4, the results on the ZHIHU dataset demonstrate that the proposed GDA-MFF achieves the best performance (ACC= 86.40%, F1 = 85.24%), significantly outperforming all ablation variants. Removing the LLM-driven module leads to substantial degradation, highlighting the critical role of generative augmentation in alleviating data sparsity and enhancing semantic coverage. Moreover, the comparison between EV-D+MFF and CGST-D+MFF shows that clustering-guided semantic template generation is essential for preserving semantic consistency, as its removal results in a notable 11.92% drop in F1 score. Overall, these results indicate that the synergy between generative augmentation and multidimensional feature fusion is crucial for effective psychological text classification.

Table 4: Ablation results of different components (%).

Model	SWM		MR		ZHIHU	
	Acc	F1	Acc	F1	Acc	F1
RF	41.12	36.20	54.06	47.93	50.10	33.19
EV-D+MFF	42.31	41.35	70.32	69.16	79.81	72.00
CGST-D+MFF	47.96	47.16	71.21	70.94	84.36	83.92
MFF	58.12	47.36	67.28	66.95	62.05	56.75
GDA-MFF	**94.80**	**94.40**	**95.90**	**95.70**	**88.67**	**86.24**

4.5 Parameter Sensitivity Analysis

In our algorithm, only two parameters for HDBSCAN, neighborhood size min_samples for automatic prompt template generation and K-means, cluster number H for feature fusion, are required. The sensitivity for these two parameters on our algorithm is analyzed here.

HDBSCAN min_samples Experiments are conducted with min_samples $\in \{3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20\}$ on 1024-dimensional BERT embeddings. As shown in Table 5, when min_samples $= 3$, the noise ratio is lowest (59.4%), the cluster count is 3, and the DBCV score is highest. This result indicates that, in the small-sample and high-noise setting of psychological texts, a smaller min_samples value more effectively preserves fine-grained psychological semantic clusters. It prevents excessively large parameter values from labeling too many samples as noise or inducing over-merging between clusters, thereby preserving implicit associations and subtle differences in psychological states. Consequently, min_samples $= 3$ is selected for all experiments above.

K-Means cluster number H Experiments are conducted with $H \in \{2, 3, 4, 5\}$. Results in Figure 7 indicate that $H = 4$ yields stable model performance across all three datasets, achieving an effective balance between representational richness and generalization.This can be attributed to the inherent characteristics of psychological texts, which include implicit semantic expressions, fine-grained emotional variations, and the coexistence of multiple potential psychological states within a single category.The $H = 4$ configuration enables an overcomplete partitioning of the feature space, allowing the clustering process to uncover latent substructures within each category and thereby enhancing the representational capacity of the fused features.However, when H is excessively large, it may introduce redundant subspaces and amplify noise, thereby increasing model variance and the risk of overfitting.

Table 5: Results of different HDBSCAN `min_samples` values.

`min_samples`	Clusters	Noise (%)	DBCV
3	3	59.4	0.002
4	2	60.2	0.000
5	2	62.4	0.001
6	2	67.7	0.001
7	2	73.7	0.001
8	2	75.9	0.001
9	2	79.7	0.001
10	0	100.0	0.000
12	0	100.0	0.000
15	0	100.0	0.000
20	0	100.0	0.000

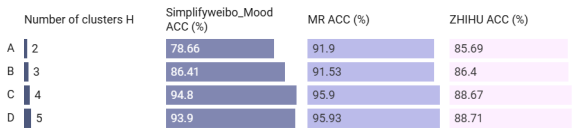

Figure 7: Classification accuracy under different K-Means cluster numbers H.

5 Conclusions

We propose a patient-oriented psychological state classification method based on LLM-driven data augmentation and multidimensional feature fusion.Despite the promising results, this study relies solely on textual information, and future research will focus on introducing multimodal fusion techniques to jointly model text with speech, images, and physiological signals, thereby enabling richer and more fine-grained psychological state recognition.

Limitations

The ZHIHU dataset, while strictly anonymized, is relatively small in scale. Generalizability to other languages, clinical settings, and disease types remains to be validated in future work.

References

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

Mohamed Bahgat, Steve Wilson, and Walid Magdy. 2022. Liwc-ud: classifying online slang terms into liwc categories. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 422–432.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Ernie Chang, Xiaoyu Shen, Dawei Zhu, Vera Demberg, and Hui Su. 2021. Neural data-to-text generation with lm-based text augmentation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 758–768.

Canyu Chen and Kai Shu. 2023. Promptda: Label-guided data augmentation for prompt-based few shot learners. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 562–574.

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.

Shan Gao, Kaixian Yu, Yue Yang, Sheng Yu, Chenglong Shi, Xueqin Wang, Niansheng Tang, and Hongtu Zhu. 2025. Large language model powered knowledge graph construction for mental health exploration. *Nature Communications*, 16(1):7526.

Yanrong Guo, Jilong Liu, Lei Wang, Wei Qin, Shijie Hao, and Richang Hong. 2023. A prompt-based topic-modeling method for depression detection on low-resource data. *IEEE Transactions on Computational Social Systems*, 11(1):1430–1439.

Amelie Gyrard, Seyedali Mohammadi, Manas Gaur, and Antonio Kung. 2024. Iot-based preventive mental health using knowledge graphs and standards for better well-being. In *Smart Technologies for Sustainable Development Goals*, pages 146–170. CRC Press.

Ernst Hellinger. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference*, pages 7184–7190.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. 2025. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2155–2171.

Nhut-Thanh Le-Hinh, Hoa-Vien Vo-Hoang, Trong-Nghia Tran, and Xuan-Nam Cao. 2023. Improving depression classification in social media text with transformer ensembles. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, pages 554–561.

Lin Lin, Xuri Chen, Ying Shen, and Lin Zhang. 2020. Towards automatic depression detection: A bilstm/1d cnn-based model. *Applied Sciences*, 10(23):8701.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Jingfang Liu and Mengshi Shi. 2022. A hybrid feature selection and ensemble approach to identify depressed users in online social media. *Frontiers in Psychology*, 12:802821.

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*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual*

meeting of the association for computational linguistics (volume 1: Long papers), pages 174–184.

Neetesh Kumar Nema, Vivek Shukla, Amit Pimpalkar, and SR Tandan. 2024. sentiment analysis of depression and anxiety social media tweets using tf-idf weighting and supervised learning algorithm. In *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, pages 1–6. IEEE.

Aytug Onan, Serdar Korukoglu, and Hasan Bulut. 2016. Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, 7(1):101–119.

Muhammad Rizwan, Muhammad Faheem Mushtaq, Urooj Akram, Arif Mehmood, Imran Ashraf, and Benjamin Sahelices. 2022. Depression classification from tweets using small deep transfer learning language models. *IEEE Access*, 10:129176–129189.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Nisha P Shetty, Balachandra Muniyal, Arshia Anand, Sushant Kumar, and Sushant Prabhu. 2020. Predicting depression using deep learning and ensemble algorithms on raw twitter data. *International Journal of Electrical and Computer Engineering*, 10(4):3751.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Ilija Tavchioski, Marko Robnik-Šikonja, and Senja Pollak. 2023. Detection of depression on social networks using transformers and ensembles. *arXiv preprint arXiv:2305.05325*.

Bazen Gashaw Teferra, Alice Rueda, Hilary Pang, Richard Valenzano, Reza Samavi, Sridhar Krishnan, and Venkat Bhat. 2024. Screening for depression using natural language processing: literature review. *Interactive journal of medical research*, 13(1):e55067.

Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. Psychbert: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082. IEEE.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yuxia Zhao, Mahpirat Mamat, Alimjan Aysa, and Kurban Ubul. 2023. Knowledge-fusion-based iterative graph structure learning framework for implicit sentiment identification. *Sensors*, 23(14):6257.

Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. 2022. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4):2823–2838.