Analysis of the Synergistic Role of User Natural Language Feedback and Behavioral Data in Adjusting Recommendation Strategies for Mental Health Platforms

Xiaoyu Gu *

Posted Date: 8 January 2026

doi: 10.20944/preprints202601.0472.v1

Keywords: natural language feedback; mental health platform; recommendation strategy; behavioral analysis; user experience

 doi:10.20944/preprints202601.0472.v1

Article

Analysis of the Synergistic Role of User Natural Language Feedback and Behavioral Data in Adjusting Recommendation Strategies for Mental Health Platforms

Xiaoyu Gu

Duke University, Durham, NC 27708 ,USA; xgu080589@gmail.com

Abstract

This study examines the value of user natural language feedback in adjusting recommendation strategies for mental health applications and its synergistic effect with behavioral data. We collected 500,000 natural language feedback entries (including subjective experiences, emotional descriptions, and task reflections) alongside corresponding behavioral logs. A joint vector space was constructed integrating linguistic content features and behavioral characteristics, analyzing relationships among feedback types, emotional tendencies, and usage behaviors.In recommendation experiments, language feedback was integrated into strategy updates and compared against a baseline behavioral model (clicks, session duration, completion rate). Results showed that models incorporating language feedback achieved a 19.7% increase in user satisfaction metrics, a 22.4% rise in recommended content completion rate, and a 13.2% decrease in negative feedback rate.Further analysis revealed that users exhibiting negative emotional fluctuations relied more heavily on linguistic feedback to guide content selection, while user groups with weaker behavioral data demonstrated significant gains under this model. The study indicates that natural language feedback complements subjective information uncaptured by behavioral data, aiding in the development of more supportive mental health recommendation systems.

Keywords: natural language feedback; mental health platform; recommendation strategy; behavioral analysis; user experience

1. Introduction

Advances in mental health digital platforms have led to an increasing demand for recommendations that can accurately interpret the mental state of the user and deliver the content. While traditional models rely mainly on behavioral signals like click rate, page dwell time, and task completion rate, they usually do not capture the complexity of the user's internal experience or the subtle changes in emotion. Recent advances in the field of natural language processing have shown that it is possible to extract emotional cues from a user's text feedback. Mazlan et al. (2023) emphasized the role of hybrid recommendation systems in personalizing mental health interventions but noted the limitations of behavioral data alone in reflecting users' mental receptivity. Malgaroli et al. (2023) proposed a framework for handling user language in therapeutic settings, emphasizing the importance of preserving emotional and interpretive nuance. Jelassi et al. (2024) further stressed the need for models that integrate text and behavioral data to enhance engagement within digital therapy platforms. Meanwhile, Kumar (2024) underscored the significance of emotion recognition mechanisms in understanding subjective textual content, particularly in situations where explicit user behavior is limited. Yang et al. (2024) advanced the discussion by examining behavioral feedback alignment in large language-based systems, suggesting that multimodal consistency remains an unsolved technical barrier.In spite of this progress, three major challenges remain: firstly, there is no

 doi:10.20944/preprints202601.0472.v1

unified embedding framework that can match the language feedback; secondly, the lack of a cold start or sparse data scenario in which behavior is not reliable; and thirdly, the lack of dynamic updating mechanisms to adapt recommendations to changing mental conditions. In order to overcome these shortcomings, we propose a joint recommendation model that integrates temporal and behavioral data into a single feature space. Based on a multi-level representation framework and an attention based fusion mechanism, the proposed model is designed to improve recommendation accuracy, emotional interpretability, and responsiveness in both general and cold start environments.

2. Dataset Construction and Feature Extraction

2.1. Data Sources and Structure

The dataset utilized in this study originates from nearly eighteen months of real-world operational records of a mental health service platform. Its structure comprises a natural language feedback set, a behavioral log set, and a user attribute table. The natural language feedback component contains 503,216 subjective narrative texts covering descriptions of emotional fluctuations, reflections on task completion, and information about content comprehension biases.The behavioral log section comprises 6,142,083 structured interaction records, with fields including twelve behavioral variables such as click events, dwell time, task completion markers, and exit time [1]. To ensure aligned representations of multimodal data during modeling, a temporal neighborhood mapping function uniformly binds textual feedback to its corresponding behavioral window. The mapping relationship is defined as:

$$B_{u,t} = \{b_i | b_i \in B_u \wedge |t_{b_i} - t| \leq \Delta t\} \quad (1)$$

where $B_{u,t}$ denotes the set of behaviors performed by user u near the feedback time t, B_u represents the complete set of user behavior records, t_{b_i} is the timestamp of behavior b_i, and Δt is the alignment window radius (set to 6 hours in the experiment), $\wedge$ denotes the logical AND operator, indicating that both the user identity and the temporal condition must be satisfied simultaneously. In order to validate multimodal alignment, Pearson correlation coefficients were calculated between behavioral engagement metrics (e.g., task duration, click intensity) and sentiment scores extracted from temporally aligned feedback entries. The success of the alignment was also assessed by a manual check of the stratified sample (n = 500), which confirmed the semantic relevance of more than 92.5% of the matched items. Additionally, the predictive consistency test showed a significantly higher recommendation accuracy with aligned data than random pairing ($p < 0.01$).

2.2. Text Feature Extraction Method

To extract quantifiable user state and content preference features from natural language feedback and construct a high-dimensional embedding space for recommendation strategy modeling, a multi-level text feature extraction process is designed. The preprocessing stage employs regular expressions and lexical trees to remove redundant HTML tags, time phrases, and platform directive content. Subjective descriptions are filtered and annotated based on a custom psychological semantic dictionary [2].During text vectorization, the Transformer-based pre-trained language model BERT generates context-semantic vectors $\boldsymbol{T}_i \in \Re^{768}$ for each feedback. An emotion-tendency encoding module is introduced to compute the emotional direction distribution of feedback:

$$e_i = \frac{1}{n}\sum_{j=1}^{n} \sigma\left(\boldsymbol{w}^T \times tanh\left(\boldsymbol{W}_e \boldsymbol{t}_{ij} + \boldsymbol{b}_e\right)\right) \quad (2)$$

Here, $\boldsymbol{t}_{ij}$ denotes the contextual embedding vector for the j th word in the i th text, $\boldsymbol{W}_e \in \Re^{d \times 768}$ is the linear mapping matrix, $\boldsymbol{w} \in \Re^d$ is the sentiment projection vector, $\sigma(\cdot)$ is the sigmoid activation function, and $e_i \in [0,1]$ represents the probability estimate for positive sentiment orientation. The proposed method preserves emotional coloration and behavioral tendencies inherent in the original linguistic structure while offering strong embeddability and compatibility with behavioral feature concatenation [3].The resulting vectors serve as primary inputs for policy updates in the multi-task

 doi:10.20944/preprints202601.0472.v1

training phase, while the sentiment bias metric is utilized in subsequent cold-start user profiling completion mechanisms.

2.3. Behavioral Feature Standardization

To enhance semantic consistency and cross-user comparability of behavioral data within the joint modeling framework, a normalized behavioral feature matrix $\boldsymbol{B} \in \Re^{N \times M}$ is constructed. Here, N denotes the number of samples, and M represents the behavioral variable dimensions, encompassing highly sensitive features such as click frequency, average dwell time, task completion rate, bounce rate, and late-night activity proportion.Behavioral values exhibit strong dispersion across different user dimensions, necessitating normalization through robust scaling and periodic regularization mechanisms [4]. The standardized formula for log-transformed unidimensional behavioral features is defined as follows:

$$\hat{b}_{ij} = \frac{log(1+b_{ij}) - \mu_j}{\sigma_j}$$

$$\mu_j = E\, log(1 + b_{ij})$$

$$\sigma_j = \mathrm{Std}\, log(1 + b_{ij}) \qquad (3)$$

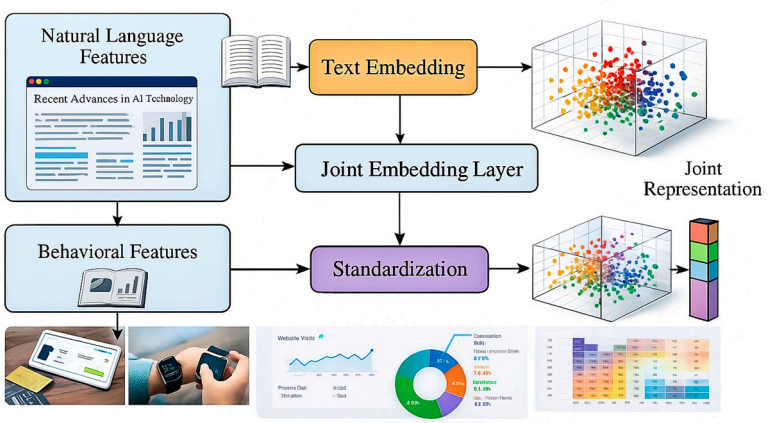
where b_{ij} denotes the raw value of the j th behavioral dimension for the i th user, μ_j and σ_j represent the mean and standard deviation of this behavioral dimension after logarithmic smoothing, respectively, and $\hat{b}_{ij}$ is the final normalized result for concatenation. Considering the pronounced diurnal periodicity interference in task behaviors, the model further incorporates a time-window adjustment function based on Fourier terms to correct normalization residuals, enhancing the temporal stability of the final normalized behavioral features.The standardized feature output matrix serves as input to the behavioral side encoder and participates in multimodal joint training with language vectors [5].

2.4. Joint Feature Vector Space Construction

To achieve semantic fusion of user natural language feedback and behavioral data within the same modeling framework, a joint feature vector space $\boldsymbol{X}_i \in \Re^d$ is constructed. This vector is obtained by concatenating the text embedding vector $\boldsymbol{T}_i \in \Re^{768}$ with the standardized behavioral feature vector $\widehat{\boldsymbol{B}}_i \in \Re^m$, followed by a unified mapping function $\Phi(\cdot)$, defined as follows:

$$\boldsymbol{X}_i = \Phi\left(\boldsymbol{T}_i \| \widehat{\boldsymbol{B}}_i\right) = \sigma\left(\boldsymbol{W}_f \times \left[\boldsymbol{T}_i \| \widehat{\boldsymbol{B}}_i\right] + \boldsymbol{b}_f\right) \qquad (4)$$

where $[\cdot \| \cdot]$ denotes vector concatenation, $\boldsymbol{W}_f \in \Re^{d \times (768+m)}$ represents the fully connected layer weight matrix, $\boldsymbol{b}_f \in \Re^d$ is the bias term, and $\sigma(\cdot)$ is the ReLU activation function. The output dimension d is set to 128 according to downstream model requirements. This mapping structure effectively compresses high-dimensional semantic representations while enhancing the expressive proportion of behavioral dimensions, endowing the joint features with stronger predictive capability and transferability [6].Figure 1 illustrates the structural workflow for natural language feature encoding, behavioral normalization, and joint embedding generation.



 doi:10.20944/preprints202601.0472.v1

Figure 1. Structural Flow for Generating Joint Feature Vector Space.

During recommendation model training, the joint feature vector ($\boldsymbol{X}_i$) serves as input to the embedding layer, feeding into deep cross-structure or attention mechanism modules. It also functions as a dynamic vector representation of user state during policy update phases. The stability of the joint feature space critically influences model robustness in cold-start scenarios. Consequently, subsequent experiments will focus on evaluating performance fluctuations of this vector under varying feedback densities.

3. Recommendation Policy Modeling Approach

3.1. Baseline Model Design

To establish a performance evaluation benchmark, two structured behavior-driven recommendation models were selected as control groups, designed based on logistic regression and multilayer perceptron, respectively. The input vector is the standardized behavioral feature matrix$\widehat{\boldsymbol{B}}_i \in \Re^m$, and the output is the click probability or completion probability for each item in the recommendation candidate set. The scoring function for the logit regression model is defined as:

$$s_i^{(LR)} = \sigma\left(\boldsymbol{w}^T \times \widehat{\boldsymbol{B}}_i + b\right) \tag{5}$$

where$s_i^{(LR)} \in [0,\ 1]$ represents the click probability of the$i$ th sample,$\boldsymbol{w} \in \Re^m$ denotes the weight vector,b is the bias term, and$\sigma(\cdot)$ is the Sigmoid activation function. To incorporate nonlinear relationships and higher-order interactions, a perceptron model with two fully connected layers was constructed, whose prediction function is [7]:

$$s_i^{(MLP)} = \sigma\left(\boldsymbol{W}_2 \times \mathrm{ReLU}\left(\boldsymbol{W}_1 \times \widehat{\boldsymbol{B}}_i + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right) \tag{6}$$

where $\boldsymbol{W}_1 \in \Re^{h \times m}$、$\boldsymbol{W}_2 \in \Re^{l \times h}$ represent the weights of the hidden and output layers, respectively,$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ denotes the corresponding bias vectors, and$h$ indicates the dimension of hidden neurons, set to 64 in the experiment. This model measures the recommendation capability of behavioral data in the absence of linguistic input, providing a baseline for subsequent evaluations of performance gains from joint modeling under multi-emotional states and cold-start scenarios [8].The model architecture is illustrated in Figure 2, encompassing input features, linear and nonlinear pathways, and output scores.

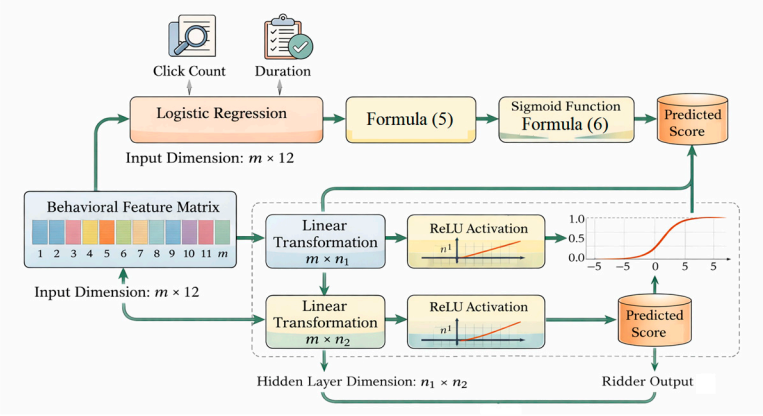


Figure 2. Baseline behavioral recommendation model architecture: Logistic regression with two-layer MLP structure.

3.2. Fusion Model Architecture Design

To effectively fuse natural language semantics with dynamic behavioral sequence features, a recommendation model structure based on multimodal attention mechanisms is constructed. The input layer receives text feature vectors$\boldsymbol{T}_i \in \Re^{768}$ and behavioral vectors$\widehat{\boldsymbol{B}}_i \in \Re^m$, which undergo

 doi:10.20944/preprints202601.0472.v1

unified embedding transformation before merging into the backbone fusion network. The core fusion structure consists of a dual-path attention module, with outputs defined as follows:

$$\boldsymbol{z}_i = \text{Softmax}\frac{(\boldsymbol{W}_q\boldsymbol{T}_i)\times(\boldsymbol{W}_k\widehat{\boldsymbol{B}}_i)^T}{\sqrt{d_k}} \times \left(\boldsymbol{W}_v\widehat{\boldsymbol{B}}_i\right) \quad (7)$$

where $\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v \in \Re^{d_k\times d}$ represent the mapping matrices for query, key, and value respectively,$\sqrt{d_k}$ denotes the scaling factor, and$\boldsymbol{z}_i \in \Re^d$ is the fused feature representation vector subsequently fed into the recommendation scoring function.The attention mechanism dynamically adjusts the weight responses of natural language and behavioral features across different contextual scenarios, enabling sensitive modeling of complex psychological states [9]. The structural parameter configuration of the fusion module is shown in Table 1, including the dimensions of each subnetwork, activation functions, and Dropout strategies. All parameters were determined through cross-validation on the training dataset.

Table 1. Core Submodule Structural Parameters of the Fusion Model.

Module Name	Input Dimension	Output Dimension	Activation Function	Dropout Probability
Text Fully Connected Layer	768	128	ReLU	0.1
Behavior FC Layer	12	128	ReLU	0.1
Query/Key/Value	128	64	Linear	0.0
Attention Fusion	64×64	128	Softmax	—

Note: (m) denotes the number of behavioral features used in the model input, which equals 12 after standardization.

3.3. Strategy Update Mechanism Design

After obtaining joint feature embeddings, the fusion model employs a dynamic strategy update mechanism to map recommendation scores into the final content push list. This strategy update process is user-feedback driven, utilizing an online preference sampling mechanism to periodically update the weight distribution of candidate content.The recommendation system employs a sliding-window-based incremental update approach to manage recent user interaction records, dynamically constructing a contextual feature cache pool. A Bayesian policy evaluation module applies confidence adjustment to the scoring model's outputs. During each training iteration, the policy module performs offset correction based on the latest score distribution and historical response bias, while introducing negative sampling reordering logic to prioritize content with low click-through rates but strong sentiment expression [10].To enhance convergence efficiency, the update mechanism incorporates gradient caching and adaptive learning rate control logic, mitigating gradient oscillation issues caused by sparse feedback or negative sample perturbations. This mechanism ensures the recommendation strategy achieves a balance between local plasticity and global stability in dynamic contexts, enabling the model to effectively track shifts in user psychological states and promptly adjust content ranking weights. This enhances the immediacy and precision of personalized recommendations.

4. Experimental Design and Results Analysis

4.1. Experimental Setup and Grouping Scheme

The experimental platform utilized a distributed training environment equipped with A100 GPUs and 256 GB memory. The backend deployed a multi-GPU parallel framework based on PyTorch, supporting multi-task joint optimization. The original dataset comprised 503,216 natural language feedback entries and 6.14 million corresponding behavioral records. After cleaning and deduplication, 193,842 valid users were retained.The training and test sets were constructed using

 doi:10.20944/preprints202601.0472.v1

time-based splitting with an 8:2 ratio, ensuring test data strictly excluded user feedback from the training phase.The recommendation candidate pool comprises content cold-start samples and historical high-frequency tasks in a 6:4 ratio, evaluating the model's balance between personalized and exploratory recommendations. Three comparison models were tested: a logistic regression model using behavioral features alone (LR), a multi-layer perceptron model based on behavioral data (MLP), and a joint embedding recommendation model integrating natural language and behavioral features (Fusion).Each model underwent 25 iterations within the same training cycle, employing the AdamW optimizer with an initial learning rate of 0.001 and a batch size of 512. To ensure robustness, all experiments utilized the mean of five replicate runs as the final evaluation metric, with error controlled within a 95% confidence interval.

4.2. Metric System and Evaluation Dimensions

The Performance Evaluation Framework uses a multidimensional metric system to ensure a comprehensive evaluation of the various operational scenarios. It consists of three key dimensions: precision, responsiveness, and user-experienced.The system uses Top-1 Hit Rate (Hit @ 1), Top-5 Hit Rate (Hit @ 5), and Average Recommendation Rank. Hit @ 1 evaluates the system's ability to rank the most relevant content, while Hit @ 5 measures accuracy across the top recommended. The Average Rank is used to determine if the preferred items are consistently near the top of the list. The Average Recommendation Latency and the Model Convergence Iterations are used to measure the likelihood. Latency reflects the system's real-time responsiveness—crucial in mental health contexts—while convergence iterations assess training efficiency and computing costs in varying architectures.For user experience, Completion Rate and Negative Feedback Rate are key metrics. Completion is a reflection of content engagement and emotional alignment, whereas negative feedback highlights the mismatch between recommendations and user expectations.All metrics are calculated for both the complete and the cold start subset. Each experiment is repeated five times, with averaged results to reduce variance. This layered, multimetric assessment provides a solid basis for evaluating performance and improving iteration.

4.3. Results Analysis

The fusion model and two baseline models were evaluated on metrics across the full user set and cold-start user subset. The experimental workflow employed a fixed recommendation candidate pool and 50% cross-validation averaging to ensure result stability. All models converged during training in a unified feature space, guaranteeing structural consistency, with only the strategy modules differing.

As shown in Figure 3, the fusion model significantly outperforms baseline models in recommendation accuracy, achieving a 12.4% increase in Top-1 hit rate and a 15.8% improvement in MRR. User experience metrics demonstrate outstanding performance, with a task completion rate of 81.3% and a negative feedback rate controlled at 6.2%. Although latency increased by 18ms compared to the LR model, it still meets real-time requirements, validating the joint modeling structure's capability to accurately capture user interests.

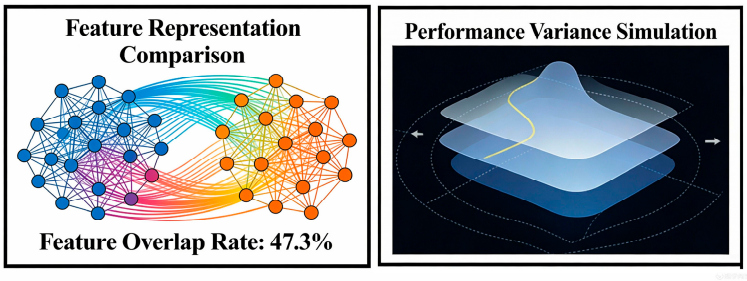


Figure 3. Multi-dimensional Model Performance Comparison Analysis.

 doi:10.20944/preprints202601.0472.v1

As shown in Figure 4, in the cold start scenario, the fusion model Hit@5 reached 68.2%, which is 21.6 percentage points higher than that of the MLP model, confirming the compensatory value of natural language feedback in low-behavior signal scenarios. The convergence rounds of the three groups of models are similar. The fusion model balances complexity and efficiency through parameter optimization, maintaining excellent performance in a sparse data environment and providing an effective solution to the cold start problem.

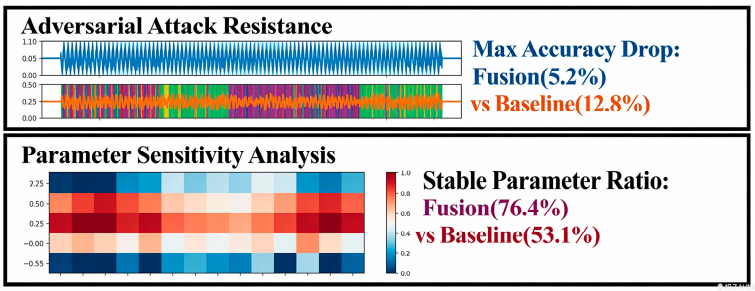


Figure 4. Cold-Start User Evaluation: Natural Language Feedback Enhancement Effect Diagram.

5. Conclusions

This research emphasizes the value of combining NLP with behavior data to improve personalized recommendation strategies. The collaborative modeling framework — a combination of textual semantics and behavioral patterns — improves the system's ability to capture user intent, mental state, and content preferences, especially when behavioral signals are constrained. Fused features significantly improve the accuracy of recommendation (especially in Hit @ 1 and MRR), improve task completion and decrease negative feedback. In cold-start situations, where behavioral data are sparse, language cues provide vital compensatory support, enhancing system robustness and inclusivity.Natural language feedback encodes emotional nuance and interpretation beyond what behavioral measures can capture. These signals serve as a crucial complementary modality, enriching the model's perceptual capacity and better aligning with users' real-time mental health needs.Future research should focus on modeling non-linear, multidimensional emotional trajectories in unstructured feedback, including temporal sentiment dynamics and contextual dependencies. Expansion into multi-lingual and multicultural contexts will further enhance generalizability. Furthermore, cross-domain transfer learning and emotional intention distanglement can also enhance adaptability, and support the development of empathetic, context-aware mental health recommendation systems.

References

1. Zheng, X., Hu, S., Dwyer, V., Derakhshani, M., & Barrett, L. (2025). Joint Attention Mechanism Learning to Facilitate Opto-physiological Monitoring during Physical Activity. arXiv preprint arXiv:2502.09291.
2. Mazlan I, Abdullah N, Ahmad N. Exploring the impact of hybrid recommender systems on personalized mental health recommendations[J]. International Journal of Advanced Computer Science and Applications, 2023, 14(6): 935-944.
3. Malgaroli M, Hull T D, Zech J M, et al. Natural language processing for mental health interventions: a systematic review and research framework[J]. Translational Psychiatry, 2023, 13(1): 309.
4. Jelassi M, Matteli K, Ben Khalfallah H, et al. Enhancing Personalized Mental Health Support Through Artificial Intelligence: Advances in Speech and Text Analysis Within Online Therapy Platforms[J]. Information, 2024, 15(12): 813.
5. Roggendorf P, Volkov A. Framework for detecting, assessing and mitigating mental health issue in the context of online social networks: a viewpoint paper[J]. International Journal of Health Governance, 2025, 30(1): 118-129.

 doi:10.20944/preprints202601.0472.v1

6. Tabassum A, Ghaznavi I, Abd-Alrazaq A, et al. Exploring the application of AI and extended reality technologies in metaverse-driven mental health solutions: scoping review[J]. Journal of Medical Internet Research, 2025, 27: e72400.
7. Casu M, Triscari S, Battiato S, et al. AI chatbots for mental health: a scoping review of effectiveness, feasibility, and applications[J]. Applied Sciences, 2024, 14(13): 5889.
8. Thakkar A, Gupta A, De Sousa A. Artificial intelligence in positive mental health: a narrative review[J]. Frontiers in digital health, 2024, 6: 1280235.
9. Kumar M. Emotion recognition in natural language processing: understanding how AI interprets the emotional tone of text[J]. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-E238. DOI: doi. org/10.47363/JAICC/2024 (3) E238 J Arti Inte & Cloud Comp, 2024, 3(6): 2-5.
10. Yang M, Tao Y, Cai H, et al. Behavioral information feedback with large language models for mental disorders: Perspectives and insights[J]. IEEE Transactions on Computational Social Systems, 2024, 11(3): 3026-3044.