ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 12 August 2025; Accepted: 28 December 2025

https://doi.org/10.17559/TV-20251208003192
Original scientific paper

# A Multi-Scale Deep Learning Architecture for Psychological State Recognition and Early Risk Warning from Social Media Text

Yufei CHEN, Kai CHEN*

**Abstract:** Social media has become an important channel for expressing emotional experiences and potential psychological distress, making automated psychological state recognition a key technical challenge for early risk warning systems. Psychological signals in text are distributed across multiple linguistic levels, ranging from character-level expressive variations to word-level semantics and sentence-level psychological structure, which limits the effectiveness of single-scale models. This paper proposes a multi-scale deep learning architecture for psychological state recognition from social media text. The approach integrates character-level and word-level representations, multi-scale convolutional modules for local semantic extraction, attention-based global semantic modeling, and cross-scale feature fusion. By jointly capturing fine-grained linguistic cues and global psychological context, the proposed model enhances the discriminative power of psychological representations. Experiments conducted on a multi-class mental health text dataset demonstrate that the proposed method consistently outperforms traditional machine learning models, conventional deep learning architectures, and attention-enhanced baselines in terms of accuracy, precision, recall, and F1-score. Furthermore, the model outputs are transformed into temporal risk signals, enabling the identification of weak, accumulating, and accelerating psychological risk patterns. The results indicate that multi-scale text modeling provides an effective technical solution for psychological state recognition and establishes a practical basis for the development of early psychological risk warning systems.

**Keywords:** computational mental health; early risk warning; multi-scale modelling; psychological state recognition; social media analysis

## 1 INTRODUCTION

As social media has evolved into a dominant medium for daily communication, it now plays a central role in how individuals express themselves and interact with others [1, 2]. Beyond personal communication, social media platforms have also become a rich source of behavioral and semantic data across diverse domains, enabling the analysis of user preferences, attitudes, and decision-making processes [3, 4]. These platforms have lowered the barriers to sharing personal thoughts and emotions, making self-disclosure more frequent and more immediate than ever before.As a result, an increasing number of users openly post about their emotional experiences, psychological distress, and fluctuations in mental states [5].

These textual expressions provide valuable data for automated mental health analysis, making the identification of potential psychological states from natural language an important research direction with broad application value [6]. However, mental health-related texts typically exhibit characteristics such as implicit expression, loose structure, and strong semantic ambiguity. Unlike traditional sentiment classification, psychological signals are often distributed across multiple layers of linguistic structure, which introduces substantial challenges for automatic identification [7]. How to accurately model the multi-layered semantics of psychological texts within a deep learning framework remains a key issue in current research.

Most existing studies rely on single-scale text modeling methods, such as word-level sequence modeling or sentence-level global representation learning. While such approaches may suffice in many NLP tasks [8, 9], they are inadequate for mental health text analysis. Psychological cues often exist simultaneously at multiple linguistic granularities: subtle expressive variations at the character level, semantic selection at the word level, and emotional structure at the sentence level. Psychological states are not determined by any single dimension but are shaped by the joint influence of multiple layers of linguistic information. Single-scale models are unable to establish effective connections across these layers, often overlooking fine-grained information or failing to capture the overall semantic structure, thus limiting performance in mental state recognition.

From the perspective of linguistics and psychological expression mechanisms, mental states exhibit inherent hierarchical characteristics [10]. Linguistic meaning is constructed through multiple structural components, and information at different levels serves distinct expressive functions, this phenomenon is especially evident in psychological texts. Character-level signals often reflect subtle emotional intensity and stylistic variations; word-level units convey core semantic components; and phrase- and sentence-level structures determine the overall psychological orientation. Psychological expression typically involves implicitness, complexity, and cross-level features, making it difficult for single-scale models to fully capture [11, 12]. Multi-scale modeling more closely aligns with the hierarchical nature of natural language, enabling the extraction of complementary information across different granularities. Such a design provides richer semantic perspectives and enhances the model's sensitivity to latent psychological cues.

Moreover, psychological health categories often exhibit semantic similarity and fuzzy boundaries. Mental states may show continuous variation or overlapping distributions in semantic space, which demands stronger representational discrimination from the model. Single-scale representations are limited by their expressive capacity at a specific linguistic level, whereas multi-scale representations strengthen category separation by integrating information from multiple embedding spaces. This helps build a more stable feature space and improves robustness when handling stylistically diverse or noisy psychological texts.

From a deep learning architectural viewpoint, multi-scale modeling enables hierarchical text representation through different receptive fields, encoding mechanisms,

and feature fusion strategies. Character-level, word-level, and sentence-level models each emphasize different linguistic aspects, while a multi-scale framework integrates these complementary perspectives into a unified semantic space, enabling continuous modeling from local patterns to global semantics. By establishing cross-scale associations within the model, the approach strengthens its ability to capture psychological expressions and allows more reliable inference even when facing implicit linguistic features [13]. This design fundamentally aligns with the structural characteristics of psychological text and provides a more comprehensive technical foundation for mental state recognition.

Psychological states expressed in social media text exhibit several unique challenges that complicate automatic recognition (1). Multi-layered linguistic signals: psychological cues emerge simultaneously at character, word, and sentence levels, making single-scale models insufficient to capture subtle emotional intensity, semantic nuances, and higher-level psychological logic [14](2). Implicit and heterogeneous expression: users often convey distress indirectly, through stylistic variations, non-standard writing, or scattered semantic clues, which are difficult for traditional models to interpret [15, 16](3). Fuzzy category boundaries: psychological states such as stress, depression, and anxiety frequently overlap in linguistic expression, requiring models to learn highly discriminative yet robust representations [17](4). Non-uniform semantic spans: psychological signals may appear in short emotional bursts or long descriptive segments, demanding a modeling framework capable of capturing features across diverse receptive fields.

To address these challenges, we propose a multi-scale modeling framework that integrates complementary linguistic information across different granularities. The model constructs parallel character-level and word-level representations, applies multi-scale convolutional kernels to capture local semantic patterns of varying spans, and employs an attention-based global encoding module to extract sentence-level psychological structure. A cross-scale fusion mechanism then combines fine-grained expressive cues with broader semantic context to form a unified representation optimized for psychological state recognition. This design aims to reflect the hierarchical nature of psychological expression and enhance the model's ability to identify subtle, implicit, and multi-level psychological signals in social media text.

This study makes two primary methodological contributions: (1) A multi-scale text representation framework is proposed to jointly model character-level, word-level, and sentence-level linguistic features within a unified architecture. This framework enhances the model's ability to capture implicit psychological signals both theoretically and empirically (2). A multi-scale fusion architecture is developed, which strengthens the discriminability between psychological categories through cross-granularity feature interaction and semantic integration, improving robustness and classification performance in complex linguistic environments.

Overall, this work aims to construct a mental state recognition method that fully exploits the multi-layered linguistic structure of psychological texts. The proposed approach provides a more accurate and comprehensive modeling paradigm and establishes a foundation for future research in mental health computation based on deep semantic understanding.

## 2 RELATED WORKS

### 2.1 Text-Based Analysis of Stress, Depression and Other Emotional States

With the widespread use of social media and online communication platforms, individuals increasingly articulate their daily experiences, emotional fluctuations, learning pressure, and psychological states through text [18, 19]. The openness and spontaneity of textual expression allow users to reveal psychological cues in an unintrusive manner [20]. Based on such natural language data, researchers have developed a broad range of automated text-based emotion analysis methods to capture negative affective tendencies, assess emotional polarity, and interpret semantic and linguistic patterns associated with psychological conditions. Consequently, text-based analysis has become a foundational approach in monitoring stress, depression, and related emotional states, offering a low-cost, scalable, and semantically rich source for understanding mental health risks.

In terms of methodological development, text-based psychological state detection has evolved from traditional machine-learning approaches to deep neural models capable of handling unstructured language and modeling complex semantic dependencies. Wan et al. [21] constructed a stress-detection model using social media posts and demonstrated that attention mechanisms can effectively highlight key emotional expressions relevant to stress prediction. In the educational context, Liu et al. [22] applied GCN-LSTM architectures to online learning texts, modeling inter-label correlations and temporal linguistic patterns to achieve fine-grained multi-label emotion prediction. Sankar et al. [23] conducted sentiment analysis on Twitter data for depression detection and emphasized that negative linguistic cues are strongly correlated with depressive tendencies. Collectively, these studies show that textual data can reveal not only emotional polarity but also more nuanced psychological states.

As research progresses, scholars increasingly recognize that although text provides meaningful semantic and cognitive cues, it may still be insufficient for comprehensive detection of stress and depression due to inherent ambiguities and information omissions. Consequently, more recent work incorporates text as a key component within multimodal frameworks. Tao et al. [24] integrated text, speech, and video in a multimodal spatio-temporal attention model, enabling textual semantics to complement non-verbal behavioral signals in depression detection. Similarly, Fan et al. [25] proposed a multimodal feature-enhancement network combining text with video, audio, and rPPG signals, achieving improved identification of depressive features. These developments indicate that text serves both as an effective standalone modality and as an essential semantic layer within multimodal mental-health detection systems, underscoring its indispensable role in stress and depression analysis.

### 2.2 Natural Language Processing Methods for Sentiment Analysis

Methods in natural language processing (NLP) for sentiment analysis revolve around extracting reliable emotional and attitudinal signals from textual data, aiming to map unstructured language into computational semantic representations. As the field has evolved, methodological developments have progressed from traditional bag-of-words and statistical features to deep representation learning, graph-based structures, attention mechanisms, and multi-view semantic fusion. These approaches not only focus on identifying sentiment polarity but also emphasize capturing contextual dependencies, semantic relations, knowledge associations, and multimodal symbols such as emojis. Consequently, modern sentiment analysis methods strive for a more comprehensive understanding of users' emotional expressions.

In terms of feature representation and deep semantic modeling, researchers have explored approaches that capture fine-grained opinions and sentiment cues. Yang et al. [26] proposed a sentiment-knowledge-enhanced graph attention network that integrates external affective knowledge into graph structures, enabling the model to account for both lexical dependencies and emotional semantics in aspect-based sentiment analysis. Tang et al. [27] introduced a joint modeling framework that combines sentiment analysis with topic structure learning, allowing emotional features and latent thematic representations to be learned simultaneously for a more holistic understanding of text sentiment. Additionally, Ben Ayed et al. [28] developed a machine learning model designed to interpret the semantic meaning of emoji sequences, expanding the feature space of sentiment analysis to encompass symbolic and visual forms of emotional expression commonly found in social media.

From a system architecture and large-scale processing perspective, research has increasingly shifted from static sentiment classification to real-time and high-throughput analytical pipelines capable of handling massive volumes of social media text. Gumelar et al. [29] employed an NLP-based sentiment classification workflow, including preprocessing, feature extraction, and model-based classification, demonstrating a typical system pipeline for public opinion sentiment analysis. Building on this, Ismail et al. [30] proposed a stream ETL framework for Twitter-based sentiment analysis that incorporates big data technologies to perform real-time text ingestion, cleaning, vectorization, and classification. These studies illustrate that NLP sentiment analysis is evolving from stand-alone models into comprehensive system-level methodologies encompassing feature learning, semantic fusion, graph-based modeling, and streaming computation.

## 3 METHODS

### 3.1 Overall Framework

The proposed multi-scale mental state recognition model aims to construct a textual representation capable of leveraging linguistic features across multiple levels, thereby improving the understanding of psychological expressions. The overall framework consists of four major components: multi-granularity input representation, multi-scale semantic feature extraction, global semantic modeling, and cross-scale feature fusion. These components are closely interconnected to accomplish the complete pipeline from raw text to mental state prediction.

The model begins with a multi-granularity representation of the input text, using two parallel channels at the character and word levels. The character-level input captures subtle variations in linguistic form, non-standard expressions, and weak emotional cues, while the word-level input provides core semantic information. This dual-channel design ensures that the model covers a broad range of linguistic information from the outset.

Next, the model constructs multi-scale semantic encoding structures for both character and word channels. Through multi-scale convolution or equivalent local semantic extraction mechanisms, the model learns semantic features under receptive fields of different sizes. This design enables the capture of linguistic patterns ranging from fine-grained to phrase-level spans, enhancing representational flexibility and sensitivity to diverse psychological signals.

After extracting local features, the model incorporates a sentence-level global semantic representation module to capture dependencies and the overall psychological structure of the text. Global semantic modeling complements local features by providing a holistic view of emotional dynamics and psychological states at the full-sentence level.

Finally, character-level, word-level, and sentence-level features are fused to form a unified textual representation. A classifier then predicts the psychological state based on the fused representation. By integrating multiple linguistic layers, the model constructs a more discriminative semantic space, achieving more accurate mental state classification.

### 3.2 Multi-Granularity Input Representation

To explicitly model linguistic features at different levels, the proposed framework constructs both character-level and word-level input sequences for each text instance. Given a raw text string $T$ preprocessing and segmentation yield a character sequence and a word sequence:

$$C = \left(c_1, c_2, ..., c_{L_c}\right),\ W = \left(w_1, w_2, ..., w_{L_w}\right) \tag{1}$$

where $L_c$ and $L_w$ denote the lengths of the character and word sequences, respectively.

To embed these sequences into a continuous vector space, we define two vocabularies, $v_c$ and $v_w$, along with corresponding embedding matrices:

$$E_c \in \mathbb{R}^{|\mathcal{V}_c| \cdot d_c},\ E_w \in \mathbb{R}^{|\mathcal{V}_w| \cdot d_w} \tag{2}$$

where $d_c$ and $d_w$ represent the embedding dimensions of characters and words.

For a character $c_i$ at position $i$, its embedding is obtained by indexing into the embedding matrix:

$$\boldsymbol{x}_i^{(c)} = E_c\left[\mathrm{idx}\left(c_i\right)\right] \in \mathbb{R}^{d_c},\ i = 1, ..., L_c \tag{3}$$

Similarly, the embedding for a word $w_j$ is defined as:

$$\boldsymbol{x}_j^{(w)} = E_w\left[\text{idx}\left(w_j\right)\right] \in \mathbb{R}^{d_w}, j = 1, ..., L_w \tag{4}$$

Stacking all embedding vectors produces the embedding matrices:

$$X^{(c)} = \begin{bmatrix} \left(\boldsymbol{x}_1^{(c)}\right)^\top \\ \vdots \\ \left(\boldsymbol{x}_{L_c}^{(c)}\right)^\top \end{bmatrix} \in \mathbb{R}^{L_c \cdot d_c}, X^{(w)} = \begin{bmatrix} \left(\boldsymbol{x}_1^{(w)}\right)^\top \\ \vdots \\ \left(\boldsymbol{x}_{L_w}^{(w)}\right)^\top \end{bmatrix} \in \mathbb{R}^{L_w \cdot d_w} \tag{5}$$

During batch training, sequences are padded or truncated to fixed maximum lengths $L_c^{\max}$ and $L_w^{\max}$, producing:

$$X^{(c)} \in \mathbb{R}^{L_c^{\max} \cdot d_c}, X^{(w)} \in \mathbb{R}^{L_w^{\max} \cdot d_w} \tag{6}$$

To ensure consistent treatment of features across linguistic granularities, both character-level and word-level embeddings are mapped into a shared semantic space of dimension $d$ through learnable linear transformations:

$$X^{(c)} = X^{(c)} W_c + b_c, \tilde{X}^{(w)} = X^{(w)} W_w + b_w \tag{7}$$

where $W_c \in \mathbb{R}^{d_c \cdot d}$, $W_w \in \mathbb{R}^{d_w \cdot d}$, and $b_c, b_w \in \mathbb{R}^d$ are trainable parameters.

Both transformation matrices project their original embedding spaces into the same output dimensionality, allowing subsequent multi-scale convolutions to operate on aligned feature representations. The inclusion of bias terms further stabilizes the transformation, enabling flexible adjustment of embedding distributions rather than enforcing strict linear scaling.

Although character and word sequences naturally differ in length, this unified projection does not introduce inconsistency in multi-scale feature extraction. Max-pooling over each convolutional map normalizes span differences by producing fixed-size representations, ensuring that variations in linguistic unit length do not propagate into the encoded feature space. This provides a coherent foundation for integrating fine-grained expressive cues with lexical semantic structures in later stages.

The resulting transformed inputs are:

$$\tilde{X}^{(c)} \in \mathbb{R}^{L_c^{\max} \cdot d}, \tilde{X}^{(w)} \in \mathbb{R}^{L_w^{\max} \cdot d} \tag{8}$$

These multi-granularity input representations provide a unified and comparable feature basis for the subsequent multi-scale semantic encoding stages, enabling the model to jointly process and integrate character-level and word-level information.

### 3.3 Multi-Scale Semantic Feature Extraction

To capture semantic patterns that emerge at different linguistic spans, the model applies a multi-scale encoding mechanism to both the character-level and word-level representations. Given the unified character embedding matrix $\tilde{X}^{(c)} \in \mathbb{R}^{L_c^{\max} \cdot d}$, a set of one-dimensional convolutional filters with kernel sizes $k \in \mathcal{K}$ is used to extract local features:

$$H_k^{(c)} = \text{Conv}_k^{(c)}\left(\tilde{X}^{(c)}\right) \tag{9}$$

Each convolutional operation produces a feature map $H_k^{(c)} \in \mathbb{R}^{L_c^{\max} \cdot d_k}$, reflecting character-level semantic patterns captured under receptive field $k$. To obtain fixed-dimensional representations, max-pooling is applied to each feature map:

$$g_k^{(c)} = \text{maxpool}\left(H_k^{(c)}\right) \tag{10}$$

The pooled vectors from all kernel sizes are then concatenated to form the character-level multi-scale local representation:

$$h_{\text{local}}^{(c)} = \left[g_{k_1}^{(c)}; g_{k_2}^{(c)}; ...; g_{k_m}^{(c)}\right] \tag{11}$$

This multi-scale design allows the character channel to capture psychological cues expressed in fine-grained writing behaviors. Small kernels (e.g., $k$ = 1) detect intensity-related phenomena such as character repetition, elongation, or stylistic variations, while medium kernels capture short emotional fragments, and larger kernels aggregate slightly longer expressive patterns. These phenomena are commonly associated with spontaneous emotional states in social media text and cannot be reliably captured at the word level.

In a parallel manner, the same multi-scale encoding mechanism is applied to the word-level embedding matrix $\tilde{X}^{(w)} \in \mathbb{R}^{L_w^{\max} \cdot d}$. Convolutional filters with identical kernel scales extract localized semantic patterns at the lexical and phrasal levels:

$$H_k^{(w)} = \text{Conv}_k^{(w)}\left(\tilde{X}^{(w)}\right) \tag{12}$$

followed by max-pooling:

$$g_k^{(w)} = \text{maxpool}\left(H_k^{(w)}\right) \tag{13}$$

and concatenation:

$$h_{\text{local}}^{(w)} = \left[g_{k_1}^{(w)}; g_{k_2}^{(w)}; ...; g_{k_m}^{(w)}\right] \tag{14}$$

Unlike the character channel, the word-level convolution extracts psychologically meaningful semantic

compositions, such as emotion words ("tired", "hopeless"), symptom descriptions, or short phrasal patterns ("can't sleep", "feel empty"). Different kernel sizes correspond to increasingly broader semantic constructs, from individual emotional tokens to multi-word expressions that often signal specific psychological tendencies.

Through this design, the model obtains two complementary multi-scale representations that jointly encode fine-grained stylistic cues from the character level and psychologically relevant semantic compositions from the word level. These representations, denoted collectively as

$$h_{\text{local}} = \left\{ h_{\text{local}}^{(c)}, h_{\text{local}}^{(w)} \right\} \tag{15}$$

constitute the set of local semantic features that serve as essential inputs to the subsequent global semantic modeling and cross-scale fusion processes.

By combining these two granularities, the model can jointly account for subtle emotional intensifiers at the character level and structured psychological semantics at the word level. This multi-source local representation forms a robust basis for later global attention, enabling the system to detect not only specific emotional cues but also the broader psychological narrative trajectory of the text.

### 3.4 Global Semantic Representation

While multi-scale convolutions capture localized semantic patterns, psychological states expressed in text often depend on broader contextual structures that extend beyond short linguistic fragments. To model such dependencies and obtain a holistic representation of the entire sequence, the model incorporates a global semantic encoding mechanism based on attention pooling. Given the word-level embedding sequence $\tilde{X}^{(w)} = \left( \tilde{X}_1^{(w)}, ..., \tilde{X}_{L_w^{\max}}^{(w)} \right)$, a trainable query vector $q$ is used to compute attention weights over all positions. For each position $i$, the weight is computed as:

$$\alpha_i = \frac{\exp\left( q^\top W_a \tilde{X}_i^{(w)} \right)}{\sum_{j=1}^{L_w^{\max}} \exp\left( q^\top W_a \tilde{X}_j^{(w)} \right)} \tag{16}$$

where $W_a$ is a learnable projection matrix that aligns the query with the token representations. The normalized coefficients $\alpha_i$ reflect the relative importance of each token for constructing the global semantic profile of the text.

Using these weights, the global semantic representation is obtained as a weighted sum of all word-level embeddings:

$$h_{\text{global}} = \sum_{i=1}^{L_w^{\max}} \alpha_i \tilde{X}_i^{(w)} \tag{17}$$

This mechanism aggregates contextual information across the entire sequence, allowing the model to capture semantic progression, patterns, and overarching psychological orientation. Compared with heavier architectures such as full self-attention or recurrent networks, attention pooling remains computationally efficient while still offering interpretability by highlighting psychologically salient tokens. Such characteristics make it particularly suitable for mental-health, related text, where key signals are often sparse, indirect, and distributed across different parts of a sentence.

By combining this global representation with the previously extracted multi-scale local features, the model integrates both fine-grained linguistic cues and higher-order semantic organization, forming a comprehensive basis for the subsequent fusion and classification stages.

### 3.5 Final Fusion, Classification, and Training Objective

After obtaining both the multi-scale local semantic features and the global semantic representation, the model integrates these heterogeneous sources of information into a unified representation suitable for final prediction. The fusion process begins by concatenating the character-level local representation $h_{\text{local}}^{(c)}$, the word-level local representation $h_{\text{local}}^{(w)}$, and the global representation $h_{\text{global}}$. This yields a comprehensive multi-scale feature vector that captures fine-grained stylistic cues, mid-level lexical structures, and high-level contextual semantics. Formally, the fused representation is expressed as:

$$h = \left[ h_{\text{local}}^{(c)}; h_{\text{local}}^{(w)}; h_{\text{global}} \right] \tag{18}$$

To enhance the expressive capacity of the fused vector and allow interactions between different scales of information, the representation is further transformed through a fully-connected layer with nonlinear activation:

$$z = \sigma\left( W_f h + b_f \right) \tag{19}$$

where $W_f$ and $b_f$ are trainable parameters, and $\sigma(\cdot)$ denotes a nonlinear activation function such as ReLU. This step refines the multi-scale representation and prepares it for the classification stage.

The final prediction of the psychological state is produced by a softmax classifier applied to the transformed vector $z$:

$$\hat{y} = \text{softmax}\left( W_{\text{cls}} z + b_{\text{cls}} \right) \tag{20}$$

where $W_{\text{cls}} + b_{\text{cls}}$ and are the parameters of the classification layer, and $\hat{y}$ denotes the predicted probability distribution over the mental health categories.

During training, the model parameters are optimized by minimizing the cross-entropy loss between predicted and true labels. For a training set containing $N$ samples, the objective function is defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} y_i^\top \log\left( \hat{y}_i \right) \tag{21}$$

where $y_i$ is the one-hot ground-truth label of the $i$-th sample, and $\hat{y}_i$ is the model prediction. This loss encourages the predicted distribution to align closely with the actual mental health category, and gradients are back-propagated through all components of the embedding layers, multi-scale convolution modules, attention-based global encoding, and the fusion classifier.

Through this final fusion, classification, and optimization process, the proposed model forms a complete pipeline that transforms raw text into multi-scale representations and ultimately predicts the corresponding psychological state with enhanced accuracy and robustness.

## 4 EXPERIMENTAL RESULTS

### 4.1 Data and Experimental Setup

The experiments in this study are conducted on a mental health text corpus constructed by integrating several publicly available datasets. The corpus consists of short social media posts, forum messages, and other forms of user-generated text, each annotated with one of seven mental health categories: Normal, Depression, Suicidal, Anxiety, Stress, Bi-polar, and Personality Disorder. These categories cover a broad spectrum of psychological states commonly observed in online expressions and provide a suitable foundation for evaluating the proposed multi-class classification framework.

Prior to model training, all text samples undergo a standardized preprocessing pipeline. The procedure includes normalization of raw strings, cleaning of HTML artifacts and redundant symbols, removal of meaningless special characters, and case normalization for English content. Word segmentation is performed using a standard tokenizer to obtain word-level sequences, while character-level sequences are constructed directly from the raw text. This preprocessing ensures consistent input quality across both granularities and facilitates robust multi-scale modeling.

The dataset is randomly split into training, validation, and test sets with a ratio of 8:1:1. The training set is used to learn model parameters, the validation set is utilized for hyperparameter tuning and monitoring training dynamics, and the test set serves as the final evaluation benchmark. Class distribution is preserved across all splits to avoid sample imbalance that may adversely affect model learning.

To ensure a fair assessment of the proposed framework, all experiments are conducted under a unified experimental environment. The model is trained end-to-end, with both character-level and word-level embeddings initialized randomly and kept trainable. The embedding dimension is set to 100 or 200 depending on the model configuration. The multi-scale convolution module employs kernel sizes such as 1, 3, 5, and 7, with 100-200 channels allocated per kernel size. The attention pooling module uses a trainable query vector, and its hidden dimension matches the shared semantic space defined during embedding projection.

The model is optimized using the Adam optimizer with an initial learning rate of $1 \times 10^{-3}$, combined with a learning rate decay schedule to ensure stable convergence. The batch size is set to 32 or 64, and the number of training epochs is determined dynamically based on validation performance. To mitigate overfitting, dropout and L2 regularization are employed throughout training. All experiments are run under identical hardware and software configurations to guarantee reproducibility.

Throughout the experimental process, the same input formats and parameter settings are maintained across all runs, ensuring comparability between different modules and fusion strategies. The full experimental pipeline begins with the multi-granularity text representations, proceeds through multi-scale local encoding, global semantic modeling, and cross-scale feature fusion, and ultimately produces mental health predictions on the test set. This setup ensures that the final evaluation faithfully reflects the effectiveness of the proposed multi-scale modeling approach.

### 4.2 Comparative Experiments

To comprehensively evaluate the effectiveness of the proposed multi-scale model in mental state recognition, this section compares it with a set of representative existing methods. The experiments adopt Accuracy, Precision, Recall, and F1-score as the evaluation metrics, and all metrics are reported in their macro-averaged form (Macro) to mitigate the influence of class imbalance. For the label set $C = \{1, \ldots, K\}$, the performance for each class $c$ is defined as:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{22}$$

where $\text{TP}_c$, $\text{FP}_c$ and $\text{FN}_c$ denote the numbers of true positives, false positives, and false negatives for class $c$, respectively. The F1-score for class $c$ is computed as:

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{23}$$

The macro-averaged metrics are then defined as:

$$\text{Macro-P} = \frac{1}{K}\sum_{c=1}^{K}\text{Precision}_c \tag{24}$$

$$\text{Macro-R} = \frac{1}{K}\sum_{c=1}^{K}\text{Recall}_c \tag{25}$$

$$\text{Macro-F1} = \frac{1}{K}\sum_{c=1}^{K}\text{F1}_c \tag{26}$$

Accuracy is given by:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\hat{y}_i = y_i\right) \tag{27}$$

where $N$ denotes the number of test samples, $\hat{y}_i$ is the predicted label, and $\mathbb{I}(\cdot)$ is the indicator function.

To objectively validate the effectiveness of the multi-scale modeling strategy, the study includes a diverse range of baseline models from traditional statistical

approaches to modern deep learning and pre-trained language models. The traditional baselines consist of TF-IDF combined with a linear SVM classifier and TF-IDF combined with logistic regression. These models rely on sparse representations derived from term frequency statistics and represent the performance ceiling of shallow text classification methods.

The deep learning baselines include three representative architectures. The first is the convolutional neural network (CNN), which uses fixed-width convolutional filters to capture local n-gram features. The second is the bidirectional long short-term memory network (BiLSTM), which captures bidirectional contextual sequence information. The third is BiLSTM enhanced with an attention mechanism (BiLSTM + Attention), which assigns adaptive weights to important psychological cues within a sentence. These architectures respectively represent the typical abilities of deep models in local pattern extraction, sequential modeling, and sentence-level attention modeling.

Tab. 1 reports the performance of all models on the test set. The results show a clear upward trend in performance as model capacity and semantic modeling sophistication increase.

**Table 1** Performance comparison among baseline methods and the proposed model

| Model | Accuracy | Macro-P | Macro-R | Macro-F1 |
|---|---|---|---|---|
| SVM + TF-IDF | 70.8 | 67.4 | 63.9 | 65.2 |
| LR + TF-IDF | 69.3 | 65.8 | 63.1 | 64.4 |
| CNN | 74.6 | 71.2 | 69.4 | 70.3 |
| BiLSTM | 75.8 | 72.1 | 70.8 | 71.4 |
| BiLSTM + Attention | 77.3 | 74.0 | 72.5 | 73.2 |
| Proposed model | 86.7 | 84.3 | 83.5 | 83.9 |

Overall, traditional baselines perform the weakest, indicating that manually crafted statistical features are insufficient for capturing the subtle emotional cues and psychologically meaningful signals embedded in mental health texts. Deep learning models achieve better performance, as distributed representations enable them to encode richer semantic and contextual information. Specifically, CNN captures local lexical and stylistic patterns, BiLSTM models bidirectional contextual dependencies, and BiLSTM + Attention further enhances sentence-level semantic aggregation.

The comparative results underscore the inherently multi-scale nature of psychological expression. Models operating at a single granularity are unable to capture the diverse linguistic cues that span from fine-grained expressive details to higher-order semantic organization. By integrating multi-granularity inputs, multi-scale local semantics, and global sentence-level structures, the proposed method effectively combines complementary information sources. This comprehensive modeling capability leads to a more discriminative representation of psychological states, which explains its consistently higher macro-F1 performance compared with all baselines.

### 4.3 Ablation Study

To thoroughly examine the contribution of different scale-related components within the proposed model, this section conducts a systematic ablation study focusing exclusively on the multi-scale architecture. The goal is to evaluate how single-scale modeling, character-level multi-scale modeling, word-level multi-scale modeling, local full multi-scale modeling, and global semantic modeling individually and jointly influence performance. Mental health texts often contain fine-grained expressive signals, mid-level semantic constructions, and psychological logic, which makes multi-scale modeling essential. The ablation experiments help reveal why the full multi-scale framework significantly outperforms single-scale or partial-scale models.

All ablation variants share identical training settings, data splits, and optimization parameters to ensure that performance differences arise solely from structural modifications. Accuracy, Macro-P, Macro-R, and Macro-F1 are reported, with Macro-F1 serving as the primary indicator. The following model variants were evaluated:

(1) Single-Scale Convolution: both character and word channels use a single fixed convolution kernel to assess the baseline performance without scale variability. (2) Character-Only Multi-Scale: only the character channel adopts multi-scale convolution to evaluate the effect of fine-grained scale expansion. (3) Word-Only Multi-Scale: only the word channel uses multi-scale convolution to assess the importance of multi-scale semantic patterns at the lexical level. 4) Local Full Multi-Scale: both character and word channels use multi-scale convolution, while global semantic modeling is removed; this variant examines the upper bound of purely local multi-scale modeling. (5) Full Multi-Scale + Global Modeling: the complete model proposed in this study, combining character-level, word-level, local multi-scale, and global semantic features. The experimental results are shown in Tab. 2.

**Table 2** Ablation results of different multi-scale configurations

| Model Variant | Accuracy | Macro-P | Macro-R | Macro-F1 |
|---|---|---|---|---|
| Single-Scale | 80.2 | 77.4 | 76.1 | 76.6 |
| Character-Only Multi-Scale | 82.7 | 79.9 | 78.4 | 79.2 |
| Word-Only Multi-Scale | 84.6 | 81.8 | 80.9 | 81.3 |
| Local Full Multi-Scale | 85.1 | 82.4 | 81.6 | 82.0 |
| Proposed model | 86.7 | 84.3 | 83.5 | 83.9 |

The ablation results reveal a clear hierarchy of contributions across different modeling scales, demonstrating that mental state recognition depends on the complementary strengths of fine-grained, lexical, and global semantic features.

The single-scale convolution baseline yields the lowest performance (78.6% Macro-F1), indicating that psychological signals span heterogeneous linguistic ranges, from short expressive bursts to longer descriptive structures, making fixed receptive fields insufficient.

Introducing multi-scale modeling at the character level brings a modest improvement (79.2% Macro-F1), confirming the importance of fine-grained expressive cues such as elongation, repetition, punctuation variation, and other stylistic signals that often accompany emotional disclosure.

Multi-scale modeling at the word level offers a larger gain (81.3% Macro-F1), highlighting that lexical semantic compositions, emotion words, symptom phrases, and behavior descriptions, carry core psychological information at variable spans. Combining the two (Local Full Multi-Scale) further improves performance to 82.0% Macro-F1, showing that character- and word-level cues are highly complementary: the former captures expressive detail, while the latter provides semantic structure.

The best performance (83.9% Macro-F1) emerges when global semantic modeling is added. This confirms that local cues alone, even when modeled at multiple scales, cannot capture emotional trends or sentence-level psychological logic. The attention-based global module integrates salient information across the entire text, producing a coherent global interpretation that aligns with how psychological states are conveyed in extended narrative context.

Overall, the ablation study shows that each component contributes a distinct and indispensable capability: character-level multi-scale modeling enhances detail sensitivity, word-level multi-scale modeling strengthens semantic extraction, local full multi-scale modeling provides layered linguistic representation, and global modeling supplies holistic contextual coherence. The complete model outperforms all variants because it unifies these complementary strengths into a single, comprehensive multi-scale framework.

### 4.4 Sensitivity Analysis

To further assess the robustness and stability of the proposed multi-scale framework under varying parameter settings, a comprehensive sensitivity analysis is conducted across three major dimensions: the number of convolution channels, the configuration of convolutional kernel scales, and the weighting balance between character-level and word-level features during the fusion stage. All experiments are carried out using the full model as the base architecture, with controlled variation of a single factor while keeping the remaining components, datasets, and optimization strategies unchanged. This design ensures that observed performance differences can be attributed exclusively to the parameter being examined. Evaluation metrics include Accuracy, Macro-P, Macro-R, and Macro-F1, with Macro-F1 serving as the primary measure due to its reliability under class imbalance.

The analysis first explores the effect of varying the number of convolution channels per scale. Channel numbers are set to 50, 100, 150, and 200 while keeping all other settings constant. The results are presented in Tab. 3.

**Table 3** Sensitivity analysis of convolution channel numbers

| Channels | Accuracy | Macro-P | Macro-R | Macro-F1 |
|---|---|---|---|---|
| 50 | 84.2 | 80.5 | 79.8 | 80.1 |
| 100 | 85.9 | 83.1 | 82.4 | 82.7 |
| 150 | 86.7 | 84.3 | 83.5 | 83.9 |
| 200 | 86.5 | 84.0 | 82.9 | 83.3 |

The results indicate that insufficient channel capacity (e.g., 50 channels) limits the representational richness of the multi-scale convolutional blocks, making it difficult for the model to capture the diverse patterns present in mental health texts. As channels increase to 100-150, performance improves substantially and stabilizes, suggesting that this range provides adequate expressive power without excessive redundancy. However, increasing the channel count to 200 does not result in further gains and instead introduces mild overfitting. This indicates that while the model requires a certain level of feature capacity to encode multi-scale patterns effectively, its performance does not depend on extremely large channel sizes, demonstrating robust behavior across a reasonably wide parameter range.

The second part of the analysis evaluates the sensitivity of the model to different combinations of convolution kernel sizes. Mental health expressions involve highly variable spans, from short emotional bursts and symbolic repetitions to longer semantic constructs, making kernel size an essential design factor. The tested kernel configurations include a single fixed size, pairs of sizes, moderate combinations, and larger-scale sets. The results are shown in Tab. 4.

**Table 4** Sensitivity analysis of convolution kernel scale configurations

| Kernel Set | Accuracy | Macro-P | Macro-R | Macro-F1 |
|---|---|---|---|---|
| {3} | 81.1 | 77.6 | 75.9 | 77.0 |
| {3, 5} | 83.5 | 80.7 | 79.8 | 80.2 |
| {1, 3, 5} | 85.2 | 82.3 | 81.1 | 81.7 |
| {1, 3, 5, 7} | 86.7 | 84.3 | 83.5 | 83.9 |
| {1, 3, 5, 7, 9} | 86.1 | 83.6 | 82.8 | 83.2 |

The results clearly demonstrate that single-scale convolution is inadequate for modeling the diverse linguistic spans inherent in mental health texts, resulting in the lowest performance. Expanding to two kernel sizes yields noticeable improvement by enabling the model to capture both short and medium-length semantic patterns. Multi-scale configurations with three or four kernels provide the best outcomes because they align well with the natural distribution of psychological expressions, which frequently involve short symbolic cues (such as repeated characters), medium-length phrases describing symptoms or emotions, and larger structures conveying emotional progression. When kernel sizes become too large, as in configurations involving size 9, the model begins to capture excessive noise, slightly impairing its ability to focus on psychologically salient segments. These outcomes show that multi-scale design is effective not because more scales are always better, but because appropriate coverage of typical linguistic spans is crucial for robust psychological signal modeling.

The third dimension of sensitivity analysis addresses the relative importance of character-level and word-level representations during the fusion phase. To evaluate how the balance between fine-grained expressive signals and semantic-level structures influences the final performance, different weighting ratios are assigned to the two feature sources. The tested ratios range from equal weights to heavily word-dominant configurations. The results are summarized in Tab. 5.

**Table 5** Sensitivity analysis of character-word feature weighting

| Word: Character Weight | Accuracy | Macro-P | Macro-R | Macro-F1 |
|---|---|---|---|---|
| 1:1 | 85.4 | 82.5 | 81.3 | 81.9 |
| 1.5:1 | 86.7 | 84.3 | 83.5 | 83.9 |
| 2:1 | 86.4 | 83.9 | 82.7 | 83.4 |
| 3:1 | 85.7 | 82.8 | 81.0 | 81.8 |

The results indicate that word-level features contribute most prominently to psychological state recognition, as much of the associated meaning is conveyed through explicit emotion terms, symptom-related expressions, and behavior descriptions. Character-level information, however, provides indispensable complementary cues that capture emotional intensity, orthographic variation, symbolic elongation, and other non-standard forms common in mental health discourse on social media. When character features are underweighted, the model becomes less sensitive to these fine-grained signals; when overweighted, the resulting imbalance disrupts the extraction of coherent semantic structures. The optimal ratio of 1.5:1 demonstrates that effective modeling requires a calibrated balance between detail-oriented and semantic-oriented features, ensuring that expressive subtleties and core meaning jointly support accurate classification.

The sensitivity analysis further shows that the proposed multi-scale model maintains stable performance across a broad spectrum of parameter settings. This robustness stems from three key factors:

(1) Moderate convolutional capacity, which prevents overfitting while preserving the ability to capture multi-span linguistic cues;

(2) Kernel size combinations aligned with typical psychological expression spans, allowing the model to adapt to both short emotional bursts and longer descriptive sequences;

(3) A balanced integration of character-level and word-level information, enabling the model to remain effective despite variations in expressive style.

These findings suggest that the multi-scale architecture is resilient to parameter fluctuations and adaptable to diverse linguistic conditions. Such stability reinforces its suitability for real-world mental health text analysis, where expressive patterns, stylistic variability, and message lengths differ widely across users and contexts.

## 5 DISCUSSION

This section provides an integrated discussion of the theoretical implications, model behavior, and practical relevance of the proposed framework, with particular attention to how multi-scale representations enable reliable psychological state interpretation and real-world deployment.

### 5.1 Theoretical Implications

The findings of this study offer several important theoretical implications for both psychological language processing and computational mental health research. By examining mental state recognition within a multi-scale linguistic modeling framework, the study demonstrates that psychological signals in text are not confined to a single linguistic layer, but instead emerge across multiple levels ranging from characters to words and up to the global sentence structure. This indicates that psychological expressions possess an inherent hierarchical nature, where meaning arises through the interaction of multiple linguistic units rather than through a linear accumulation of isolated symbols. Such observations align with hierarchical theories in linguistics, which emphasize that meaningful interpretation depends on the coordinated functioning of multiple structural layers. The clear advantage of multi-scale modeling reinforces the notion that psychological expression is shaped simultaneously by fine-grained, mid-level, and global linguistic features.

Furthermore, the analysis highlights the unique theoretical significance of character-level information in psychological text. Traditionally, character-level signals are often regarded as low-level or auxiliary features. Yet the current study reveals that, particularly in emotionally intense or psychologically burdened narratives, character-level variations, such as expressive elongation, repetition, orthographic distortion, and symbolic modulation, serve as meaningful carriers of psychological intensity. These signals cannot be adequately captured by word-level models alone, and thus broaden the conventional view of character-level features in natural language understanding. In the context of psychological expression, characters do not merely support tokenization or error tolerance; they function as direct reflections of emotional states and cognitive fluctuation.

The study also demonstrates that the semantic span of psychological language exhibits substantial irregularity and heterogeneity. Mental health texts often contain a mixture of brief emotional bursts, medium-length descriptive phrases, and longer narrative structures. Single-scale models, which rely on fixed receptive fields, are inherently unable to accommodate this diversity. Multi-scale convolutional structures, in contrast, better reflect the natural distribution of linguistic spans associated with psychological expression. This reveals that psychological meaning is not constructed uniformly but is distributed across diverse linguistic segments, confirming that mental state communication is a multi-span and multi-layer symbolic process.

In addition, the necessity of global semantic modeling underscores the contextual and narrative nature of psychological expression. While local features capture fine-grained emotional cues, the ultimate characterization of mental state relies on broader patterns such as emotional trajectories, semantic progression, and psychological coherence across sentences. These findings resonate with theoretical models in psycholinguistics that describe emotional narratives and psychological reasoning as structured, context-bound processes. Incorporating global semantic modeling within a multi-scale system enhances the model's ability to capture these cross-sentence dependencies, demonstrating that psychological expression is not merely a sum of local cues but a coherent, multi-layered linguistic phenomenon.

Finally, the experimental results reveal a strong complementarity among different linguistic scales rather than a hierarchical dominance of one over another. Psychological meaning is not primarily encoded at a single layer; rather, it emerges from the integration of fine-grained expressive signals, lexical semantic content,

and overarching psychological logic. The effectiveness of the multi-scale framework arises precisely because it mirrors the natural composition of psychological language: characters convey emotional nuance and intensity, words provide the semantic backbone, and sentences express broader psychological structure. This integrated view offers a robust theoretical foundation for mental health text analysis and points toward future research directions involving interpretable multi-layer psychological language models.

Taken together, the study not only validates the effectiveness of multi-scale modeling from an engineering perspective, but also reveals fundamental mechanisms underlying the organization of psychological meaning in language. These theoretical insights contribute a novel perspective to the understanding of mental health communication and establish a conceptual basis for the continued development of deep semantic models in computational mental health research.

### 5.2 Practical Implications

The proposed multi-scale modeling framework not only advances theoretical understanding of the layered structure of psychological language, but also yields several concrete implications for the development and deployment of practical mental health analysis systems. First, the demonstrated ability of the model to integrate character-level, word-level, and sentence-level information provides practical value for real-world psychological state assessment. Mental health texts collected from social media, online communities, or anonymous help-seeking platforms are often highly individualized and noisy. The results show that a multi-scale strategy is capable of maintaining strong robustness in the presence of irregular expressions, non-standard spellings, symbolic patterns, and heterogeneous linguistic styles. This indicates that the proposed approach is well suited for deployment in open-domain environments where psychological risk monitoring must operate on uncontrollable and diverse textual inputs.

Second, the finding that character-level information plays an indispensable role in psychological signal detection offers important guidance for system design. Many early warning indicators of psychological deterioration are conveyed not through explicit lexical expressions, but through subtle orthographic changes, character repetition, emotional elongation, symbolic punctuation, or other creative deviations from standard writing. Systems that rely solely on word-level or sentence-level features risk overlooking these fine-grained cues. Therefore, practical mental health computing systems should incorporate character-level modeling as a core component of input representation. Doing so enhances the ability to detect early and weak signals of emotional instability, thereby improving the responsiveness and sensitivity of risk assessment tools.

Third, the effectiveness of multi-scale convolutional structures across different semantic spans highlights their suitability for mental health monitoring applications. Psychological expression naturally involves a mixture of very short emotional bursts and longer descriptive or reflective statements. Multi-scale convolution enables the model to simultaneously capture both types of patterns, reducing the blind spots inherent in single-scale architectures. Thus, real-world systems devoted to mental state monitoring should prioritize feature extraction modules that cover multiple receptive fields so that psychological indicators embedded at different linguistic granularities can be comprehensively recognized.

In addition, the practical value of global semantic modeling is underscored by the observation that many psychological risks are not articulated explicitly in isolated fragments but emerge from broader narrative patterns. Emotional progression, shifts in motivational tone, cognitive distortions, and sustained psychological decline typically manifest at the sentence or discourse level rather than in short segments. The integration of a global attention mechanism equips practical systems with the ability to recognize these holistic structures, allowing more accurate and coherent assessment of the trajectory of an individual's psychological state. This capability extends the utility of the model beyond static classification, supporting dynamic monitoring and analysis of emotional fluctuations over time.

Finally, the sensitivity analysis demonstrates that the proposed model remains stable across a wide range of reasonable hyperparameter settings. This characteristic is particularly important for real-world deployment, where text sources vary substantially in style, length, noise level, and linguistic composition. A model that is not overly sensitive to parameter adjustments is easier to migrate across platforms and more suitable for long-term deployment with minimal maintenance costs. The robustness exhibited by the multi-scale architecture indicates that practical systems built upon this framework can maintain reliable performance under diverse and evolving conditions.

Overall, the multi-scale modeling approach proposed in this study offers a viable and effective technical pathway for implementing psychological risk detection and emotional monitoring in real-world settings. By combining fine-grained expressive cues, semantic structures, and global psychological organization into a unified representation, the model provides more accurate identification of psychological states and more sensitive detection of emotional changes. These advantages make it a promising foundation for mental health assessment, early risk intervention, and the development of intelligent psychological support systems.

### 5.3 End-to-End Application Workflow of the Multi-Scale Classification Model for Early Psychological Risk Warning

To operationalize the proposed multi-scale psychological state classification model in real-world early warning scenarios, the model must be embedded into a structured, end-to-end workflow that transforms raw user-generated text into actionable risk signals. This workflow consists of several consecutive stages, data acquisition, model inference, temporal aggregation, risk quantification, and warning activation, each connected through well-defined data structures and intermediate representations, enabling the construction of a deployable early warning pipeline (as shown in Fig. 1).

The first stage involves integrating the system with specific textual data sources such as social media platforms, online forums, anonymous Q&A services, or institution-based psychological support systems. Each incoming text entry is associated with a user identifier and timestamp. The system applies a preprocessing pipeline consistent with model training, including normalization, noise filtering, tokenization, and construction of both character-level and word-level sequences. The processed text is then passed to the multi-scale classification model for forward inference. For each text, the model outputs a psychological state label along with its probability distribution, and optionally intermediate feature vectors for downstream analysis. At this step, every text sample becomes a structured record linking "time–user–psychological state", forming the basic data unit for subsequent warning computations.

Building upon these model outputs, the system constructs user-level psychological state time series, which serve as the foundation for early risk detection. Specifically, for each user, the predicted labels and confidence scores are chronologically accumulated to form a temporal sequence. Within sliding windows (e.g., the past 24 hours, 3 days, or 7 days), the system computes the frequency of each psychological category, the mean confidence of predictions, and the evolving trend of these measures over time. A continuous risk score can then be derived using a weighted scheme, in which categories such as suicidal ideation or severe depression are assigned higher weights, while mild anxiety or general stress receive lower ones. By integrating these weighted values with text volume, category distribution, and prediction confidence, the system generates a dynamic risk curve. Window-to-window differences or rate-of-change measures can additionally be computed to indicate whether the user's psychological state is showing signs of rapid deterioration. In this manner, the classification model transitions from a "point-based predictor" to a "time-series risk signal generator", allowing the system to trace the evolution of psychological conditions from mild fluctuations to sustained decline.

Once a continuous risk signal has been established, the warning stage interprets this numerical representation into discrete warning levels and actionable responses. A set of tiered thresholds can be defined, for instance, a mild elevation of the risk score over a sustained period may trigger a "monitoring" status; simultaneous exceedance of intermediate thresholds and increased risk acceleration may indicate a "warning" status; and frequent occurrences of high-risk categories with strong confidence, along with a sharp rise in the risk curve, may signal a "high-risk" state. Each level corresponds to different system actions, such as internal logging, notifying mental health professionals, or issuing gentle, non-intrusive alerts to the user with recommendations for support. To reduce false alarms caused by isolated anomalies, the threshold structure may incorporate temporal stability requirements (e.g., conditions must be met across multiple windows) and minimum-text-volume constraints. Throughout this pipeline, the multi-scale classification model does not independently trigger warnings; rather, it serves as the core component that generates fine-grained and reliable psychological state signals, which are then embedded into temporal analysis and threshold-based decision logic, forming a comprehensive workflow from raw text to risk warning.

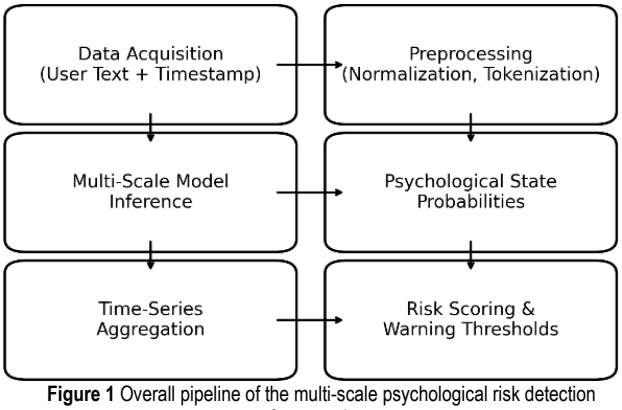


**Figure 1** Overall pipeline of the multi-scale psychological risk detection framework

## 6 CONCLUSIONS

This paper presented a multi-scale deep learning approach for psychological state recognition and early risk warning based on social media text. By integrating character-level, word-level, and sentence-level representations within a unified architecture, the proposed model effectively captures fine-grained expressive cues, lexical semantic patterns, and global psychological structure. This design overcomes the limitations of traditional single-scale text models in analyzing psychologically rich and implicitly expressed language.

Experimental results demonstrate that the proposed multi-scale framework achieves significant performance improvements over traditional machine learning methods, standard deep learning models, and attention-based baselines across multiple evaluation metrics. Ablation and sensitivity analyses further confirm the complementary roles of different linguistic scales and the robustness of the architecture under varying parameter configurations.

Beyond static classification, the study illustrates how model outputs can be aggregated over time to generate dynamic psychological risk signals, supporting early detection of weak and accumulating risk patterns. These findings suggest that multi-scale linguistic modeling provides a reliable technical foundation for developing automated psychological state recognition and early warning systems.

Future work may extend the proposed approach to multimodal behavioral data, improve interpretability, and incorporate privacy-aware mechanisms to facilitate deployment in real-world mental health monitoring applications.

## 7 REFERENCES

[1] Khadka, S. & Khadka, A. K. (2023). The role of social media in determining tourists' choices of Nepalese destinations. *Journal of Logistics, Informatics and Service Science*, *10*(3), 180-193. https://doi.org/10.33168/JLISS.2023.0314

[2] Zhu, L., De Costa, F., & Bin Yasin, M. A. (2023). Social media communication network analysis and influence propagation model: A case study. *Journal of Logistics, Informatics and Service Science*, *10*(3), 264-279. https://doi.org/10.33168/JLISS.2023.0320

[3] Shaban, A. M. (2023). The effectiveness of TV promotion and social media applications in achieving consumer brand loyalty. *Journal of System and Management Sciences*, *13*(4), 140-151. https://doi.org/10.33168/JSMS.2023.0408

[4] Laghari, A. A., He, H., Khan, A., Laghari, R. A., Yin, S., & Wang, J. (2022). Crowdsourcing platform for QoE evaluation for cloud multimedia services. *Computer Science and Information Systems*, *19*(3), 1305-1328. https://doi.org/10.2298/CSIS220322038L

[5] Boonyarat, P., Liew, D. J., & Chang, Y. C. (2024). Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19. *Information Processing & Management*, *61*(4), 103706. https://doi.org/10.1016/j.ipm.2024.103706

[6] Thekkekara, J. P., Yongchareon, S., & Liesaputra, V. (2024). An attention-based CNN-BiLSTM model for depression detection on social media text. *Expert systems with applications*, *249*, 123834. https://doi.org/10.1016/j.eswa.2024.123834

[7] Hossain, S., Umer, S., Rout, R. K., & Al Marzouqi, H. (2024). A deep quantum convolutional neural network based facial expression recognition for mental health analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 1556-1565. https://doi.org/10.1109/TNSRE.2024.3385336

[8] Mao, Y., Liu, S., & Gong, D. (2023). A hybrid technological innovation text mining, ensemble learning and risk scorecard approach for enterprise credit risk assessment. *Tehnički vjesnik*, *30*(6), 1692-1703. https://doi.org/10.17559/TV-20230316000447

[9] Cheng, Y., Wan, Y., Sima, Y., Zhang, Y., Hu, S., & Wu, S. (2022). Text detection of transformer based on deep learning algorithm. *Tehnički vjesnik*, *29*(3), 861-866. https://doi.org/10.17559/TV-20211027110610

[10] Peng, P. & Liao, Y. (2023). Six addiction components of problematic social media use in relation to depression, anxiety, and stress symptoms: a latent profile analysis and network analysis. *BMC psychiatry*, *23*(1), 321. https://doi.org/10.1186/s12888-023-04837-2

[11] Shen, X., Huang, X., Zou, S., & Gan, X. (2024). Multimodal knowledge-enhanced interactive network with mixed contrastive learning for emotion recognition in conversation. *Neurocomputing*, *582*, 127550. https://doi.org/10.1016/j.janxdis.2025.103006

[12] Li, J., Chen, N., Zhu, H., Li, G., Xu, Z., & Chen, D. (2024). Incongruity-aware multimodal physiology signals fusion for emotion recognition. Information Fusion, 105, 102220.https://doi.org/10.1016/j.inffus.2023.102220

[13] Yin, Y., Jing, L., Huang, F., Yang, G., & Wang, Z. (2024). Msa-gcn: Multiscale adaptive graph convolution network for gait emotion recognition. *Pattern Recognition*, *147*, 110117. https://doi.org/10.1016/j.patcog.2023.110117

[14] Chutia, T. & Baruah, N. (2024). A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, *57*(8), 203. https://doi.org/10.1016/j.jad.2024.08.193

[15] Khan, J., Ahmad, K., Jagatheesaperumal, S. K., & Sohn, K. A. (2025). Textual variations in social media text processing applications: challenges, solutions, and trends. *Artificial Intelligence Review*, *58*(3), 89. https://doi.org/10.1007/s10462-024-11071-z

[16] Ezerceli, Ö. & Dehkharghani, R. (2024). Mental disorder and suicidal ideation detection from social media using deep neural networks. *Journal of Computational Social Science*, *7*(3), 2277-2307. https://doi.org/10.1007/s42001-024-00307-1

[17] De la Rosa-Cáceres, A., Wendt, L. P., Zimmermann, J., & Díaz-Batanero, C. (2025). Comparing structural models for internalizing pathology: Latent dimensions, classes, or a mix of both?. *Journal of Anxiety Disorders*, *111*, 103006. https://doi.org/10.1016/j.jad.2024.08.013

[18] Chandrasekaran, R., Kotaki, S., & Nagaraja, A. H. (2024). Detecting and tracking depression through temporal topic modeling of tweets: insights from a 180-day study. *npj Mental Health Research*, *3*(1), 62. https://doi.org/10.1038/s44184-024-00107-5

[19] Kerasiotis, M., Ilias, L., & Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. *Social Network Analysis and Mining*, *14*(1), 196. https://doi.org/10.1007/s13278-024-01360-4

[20] Ahmed, O., Walsh, E. I., Dawel, A., Alateeq, K., Oyarce, D. A. E., & Cherbuin, N. (2024). Social media use, mental health and sleep: A systematic review with meta-analyses. *Journal of affective disorders*, *367*, 701-712. https://doi.org/10.1016/j.jad.2024.08.193

[21] Wan, X. & Tian, L. (2024). User stress detection using social media text: A novel machine learning approach. *International Journal of Computers Communications & Control*, *19*(5). https://doi.org/10.15837/ijccc.2024.5.6772

[22] Liu, Z., Li, F., Hao, G., He, X., & Zhang, Y. (2024). GCN-LSTM: multi-label educational emotion prediction based on graph convolutional network and long and short term memory network fusion label correlation in online social networks. *Computer Science and Information Systems*, *21*(4), 1583-1605. https://doi.org/10.2298/CSIS240314049L

[23] Sankar, P., Palanichamy, N., & Ng, K. W. (2024). Sentiment analysis on Twitter data for depression detection. *Journal of Logistics, Informatics and Service Science*, *11*(3), 21-36. https://doi.org/10.33168/JLISS.2024.0302

[24] Tao, Y., Yang, M., Li, H., Wu, Y., & Hu, B. (2024). DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 2956-2966. https://doi.org/10.1109/TKDE.2024.3350071

[25] Fan, H., Zhang, X., Xu, Y., Fang, J., Zhang, S., Zhao, X., & Yu, J. (2024). Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Information Fusion*, *104*, 102161. https://doi.org/10.1016/j.inffus.2023.102161

[26] Yang, B., Li, H., & Xing, Y. (2023). SenticGAT: Sentiment knowledge enhanced graph attention network for multi-view feature representation in aspect-based sentiment analysis. *International Journal of Computers, Communications & Control*, *18*(5), 5089. https://doi.org/10.15837/ijccc.2023.5.5089

[27] Tang, M., Cao, J., Fan, Z., Gong, D., & Xue, G. (2024). Public perceptions of EV charging infrastructure: A combined sentiment analysis and topic modeling approach. *Studies in Informatics and Control*, *33*(2), 59-72. https://doi.org/10.24846/v33i2y202406

[28] Ben Ayed, M. & Alsaawi, A. (2024). A novel machine learning model for predicting the meaning of an emojis string in social media platforms. *Studies in Informatics and Control*, *33*(1), 91-98. https://doi.org/10.24846/v33i1y202408

[29] Gumelar, G. & Girsang, A. S. (2024). Understanding public opinions of government measures against COVID-19 through Twitter sentiment analysis. *Journal of Logistics, Informatics and Service Science*, *11*(2), 1-9. https://doi.org/10.33168/JLISS.2024.0201

[30] Ismail, A., Sazali, F. H., Jawaddi, S. N. A., & Mutalib, S. (2025). Stream ETL framework for twitter-based sentiment analysis: Leveraging big data technologies. *Expert Systems with Applications*, *261*, 125523. https://doi.org/10.1016/j.eswa.2024.125523

**Contact information:**

**Yufei CHEN**
Changzhou Saixun Network Technology Co., Ltd.,

Jiangsu, China
E-mail: cyf2000404@163.com

**Kai CHEN**
(Corresponding author)
1) Changzhou No.2 People's Hospital, Jiangsu, China
2) The Third Affiliated Hospital of Nanjing Medical University,
Changzhou 213003, Jiangsu, China
E-mail: c4kaichen@163.com