Joint Prediction Model of Job Satisfaction and Turnover Tendency of College Graduates Based on Transformer-CRF

Fang Rao[1], Wei Cao[2], Jianxue Huang[3], Yi Zhou[4]
[1]Research Center for Mental Health Education, Nanchang Normal University, Nanchang 332020, Jiangxi, China
[2]Office of Admissions and Career Services,East China University of Technology, Nanchang 330013, Jiangxi, China
[3]Disciplinary Inspection and Supervision Office,Zhuhai Campus of Beijing Institute of Technology, Zhuhai 519088,Guangdong, China
[4]Department of Police Sports, Jiangxi Police College, Nanchang 330100, Jiangxi, China

Corresponding author: Fang Rao (e-mail: rf20240614@163.com)

ABSTRACT For recent college graduates, short adaptation periods and high turnover rates make joint prediction of job satisfaction and turnover tendency important for universities and employers. Existing methods usually predict satisfaction and turnover intention independently, thereby neglecting their correlation. This paper proposes a Transformer-CRF joint prediction model to capture coupled label relationships. Multi-source features, including categorical, numerical, and textual variables, are integrated through PCA and gating mechanisms. A Transformer encoder captures global dependencies among heterogeneous features, while a CRF layer models dependencies between nine satisfactionintention combination states to reduce label inconsistency. Experiments based on questionnaire data from the 2023 graduating class show that the model achieves a 96.6% perfect match rate and 98.2% joint label consistency, significantly improving prediction accuracy over baseline methods. The model also increases the 6-month retention rate by 7.2%. These findings indicate that the proposed method effectively reveals the association between job satisfaction and turnover intention and provides reliable decision support for career-stability management, early warning of employment risk, and human-resource optimization.

INDEX TERMS college graduates, transformer-crf, joint prediction, job satisfaction, label consistency

I. INTRODUCTION

AS the annual cohort of college graduates in China continues its steady increase, the overall quality of their initial employment has become a central concern for society, employers, and the graduates themselves; this is particularly pertinent in labor-intensive sectors such as the industry, where high operational demands and market volatility often influence early career experiences. Given this context, indicators such as job satisfaction and turnover tendency are crucial metrics for assessing an employee's subjective experience and an organization's stability, directly impacting long-term organizational performance and significant human resource costs [1-3].

This paper aims to construct a joint model that can simultaneously predict the job satisfaction and turnover tendency of college graduates, solving the problems of insufficient correlation modeling and difficulty in capturing long-distance dependencies between the two. This paper normalizes, embeds, and gates the heterogeneous features of graduates' categories, scales, and texts to ensure that the model can still effectively capture the long-distance dependency of features when the initial data is small. The fused features are sent to a multi-layer Transformer encoder to extract the global context representation. Finally, a CRF decoder is added to the top layer to jointly optimize the emission score and transfer matrix of the label to alleviate the label conflict problem and ensure the consistency between labels. The experiment is verified on 2512 graduate questionnaire data. The predictive model demonstrates strong performance, achieving a 96.6% joint label complete match rate, a high 95.6% prediction F1 score, and a low 1.8% label conflict rate, alongside a 0.60 MI (Mutual Information) index which quantifies the strong co-dependence captured between satisfaction and turnover.

II. RELATED WORK

At present, there have been many explorations on the prediction of job satisfaction of college graduates. Gupta A et al. used structural equation model to reveal the mechanism of action between psychological capital, psychological contract fulfillment, and organizational citizenship behavior, and further constructed a satisfaction prediction model

through random forest regression technology, verifying the value of machine learning in mining hidden features that are easily overlooked by traditional methods [4]. Based on a survey of 534 graduates from 11 Chilean universities, Espinoza O et al. used sequential logistic regression to quantitatively analyze the relationship between socio-demographic characteristics, degree matching, income level, and job satisfaction, and proposed that gender, horizontal matching, and salary level are important variables affecting satisfaction [5]. The above studies all focus on a single job satisfaction model, without simultaneously considering the turnover tendency of early-career college graduates, nor exploring the intrinsic relationship between satisfaction and turnover tendency within the same framework. Many scholars have conducted research on the prediction of the turnover tendency of college graduates and have achieved some results. Liu M et al., based on the enhanced random forest model, optimized the imbalanced sample processing and revealed key influencing factors such as income level, job satisfaction, job opportunities, and matching degree for the turnover situation of 17,268 Chinese graduates. However, they only independently predicted the turnover rate and did not consider the interaction between satisfaction and tendency [6]. Ma D and other scholars used XGBoost and SMOTE to predict employee turnover, achieving an accuracy of 87.33%, but failed to specifically target college graduates as research subjects [7]. Existing research has made many breakthroughs in algorithm innovation, feature engineering, and imbalance processing, but most of them only predict the single dimension of turnover tendency, fail to jointly model job satisfaction and turnover tendency of early-career college graduates, and fail to capture the inherent coupling relationship between the two. In addition, models such as XGBoost have a weak ability to capture long-term dependencies [8,9].

III. CONSTRUCTION OF TRANSFORMER-CRF JOINT PREDICTION MODEL

A. MULTI-SOURCE FEATURE FUSION

The experimental data in this paper is multi-source heterogeneous data of college graduates, including category features, numerical features, and text features. The three features are mapped to the same representation space, which facilitates the Transformer-CRF model to capture long-distance dependencies and label coupling relationships in parallel.

For the category vector of each sample, a learnable embedding matrix is used for mapping, as shown in Formula (1).

$$e_i = W_e x_i^{\text{cat}} + b_e \tag{1}$$

In Formula (1), ei represents the category vector; We represents the learnable embedding matrix; be represents the bias vector.

For numerical features, zero mean unit variance normalization is used, and the calculation is shown in Formula (2).

$$\hat{x}_i^{\text{num}} = \frac{x_i^{\text{num}} - \mu}{\sigma} \tag{2}$$

In Formula (2), x_i^{num} represents the numerical feature; μ represents the mean of the feature in the training set; σ represents the standard deviation.

In text feature processing, this paper directly uses TF-IDF (Term frequency-inverse document frequency) vectors, which is prone to high-dimensional sparseness. To improve computational efficiency, principal component analysis dimensionality reduction processing is used, as shown in Formula (3).

$$p_i = U^\top x_i^{\text{txt}} \tag{3}$$

In Formula (3), $U^\top$ represents the feature vector matrix of the principal components of the previous embedding dimension; x_i^{txt} represents the text feature; p_i represents the text feature after dimensionality reduction.

The three types of features are mapped to the same dimension, as shown in Formula (4).

$$\begin{cases} h_i^{\text{cat}} = W_e e_i + b_e, \\ h_i^{\text{num}} = W_n \hat{x}_i^{\text{num}} + b_n, \\ h_i^{\text{txt}} = W_t p_i + b_t. \end{cases} \tag{4}$$

In Formula (4), be, bn, and bt represent the biases corresponding to the three types of features, respectively.

To enhance the information interaction between different modalities, the final fusion vector z_i is calculated by gated fusion [10,11], as shown in Formula (5).

$$\begin{cases} u_i = \sigma(W_u[h_i^{\text{cat}}; h_i^{\text{num}}; h_i^{\text{txt}}] + b_u), \\ \hat{h}_i = \tanh(W_h[h_i^{\text{cat}}; h_i^{\text{num}}; h_i^{\text{txt}}] + b_h), \\ z_i = u_i \odot h_i^{\text{cat}} + (1 - u_i) \odot \hat{h}_i. \end{cases} \tag{5}$$

In Formula (5), σ represents element-by-element Sigmoid; $\odot$ represents element-by-element product; u_i represents gated coefficient; W_u and W_h represent weight matrices.

After fusion of features, this paper performs batch normalization on z_i, as shown in Formula (6).

$$\hat{z}_i = \gamma \frac{z_i - E_B[z]}{\sqrt{\text{Var}_B[z] + \epsilon}} + \tau \tag{6}$$

In Formula (6), EB and VarB represent the mean and variance of the current mini-batch, respectively; γ represents a learnable scaling parameter; τ represents a translation parameter; ϵ represents a small constant to prevent the value from being zero. Finally, a feature matrix Z with unified dimension and integrating multi-source information is obtained.

This paper maps multi-source features to the same dimension before inputting them into the Transformer. Text

features are first vectorized using TF-IDF to obtain a high-dimensional sparse representation, then PCA is used to reduce the dimensionality to the same dimension as the categorical and numerical features (d-dimensional embedding). Categorical features are mapped to d dimensions using a learnable embedding matrix, while numerical features are standardized and linearly mapped to d dimensions. Before gated fusion, the three types of features are superimposed according to their corresponding element positions. Learnable gating coefficients control the weights of different modal information, resulting in the final fused vector Z, which serves as the input sequence features to the Transformer.

B. TRANSFORMER MODEL

1) Position Encoding Superposition

To retain the sequence order information, this paper adds the learnable position encoding to the feature matrix. The initial input expression is shown in Formula (7).

$$X^{(0)} = Z + P \tag{7}$$

In Formula (7), $X^{(0)}$ represents the initial input, and P represents the learnable position encoding vector.

2) Multi-head Self-attention

In the Transformer model [12,13], each layer includes a multi-head attention and a feedforward network. The study first maps the output of the lth layer to query, key, and value, as shown in Formula (8).

$$\begin{cases} Q_h^{(l)} = X^{(l-1)} W_h^Q, \\ K_h^{(l)} = X^{(l-1)} W_h^K, \\ V_h^{(l)} = X^{(l-1)} W_h^V. \end{cases} \tag{8}$$

In Formula (8), $Q_h^{(l)}$ represents query; $K_h^{(l)}$ represents key; $V_h^{(l)}$ represents value. h represents the number of attention heads.

The calculation of the attention of each head is shown in Formula (9).

$$A(Q_h^{(l)}, K_h^{(l)}, V_h^{(l)}) = \mathrm{softmax}\left(\frac{Q_h^{(l)} (K_h^{(l)})^\top}{\sqrt{d_k}} \right) V_h^{(l)} \tag{9}$$

In Formula (9), $A(Q_h^{(l)}, K_h^{(l)}, V_h^{(l)})$ represents the attention output result of each head.

After the outputs of the H heads are concatenated by column, they are mapped through a linear layer, as shown in Formula (10).

$$\hat{M}^{(l)} = [\mathrm{head}_1; \ldots; \mathrm{head}_H] W^O, \quad \mathrm{head}_h = \mathrm{Attention}(\cdot) \tag{10}$$

In Formula (10), $\hat{M}^{(l)}$ represents the output after mapping.

After mapping, this paper performs residual connection and layer normalization processing, as shown in

$$S^{(l)} = \mathrm{LayerNorm}\left(X^{(l-1)} + \mathrm{Dropout}(\hat{M}^{(l)}) \right) \tag{11}$$

In Formula (11), LayerNorm represents layer normalization, and Dropout represents random inactivation.

For the multi-source feature input constructed in this paper, although its essence is structured vectors rather than traditional text sequences, the self-attention mechanism has significant advantages over standard fully connected layers. Self-attention can perform weighted interactions on all dimensions of the input vector within the same layer, capturing global dependencies between different features, while fully connected layers can only learn local linear combinations. The multi-head mechanism allows the model to focus on multiple feature subspaces in parallel, improving its ability to model potential nonlinear relationships in heterogeneous data. Finally, through residual connections and layer normalization, the self-attention module can stably train deep networks, obtaining richer joint representations when fusing categorical, numerical, and textual features. These characteristics make the Transformer architecture more expressive and generalizable than traditional fully connected layers when processing multi-source structured data.

3) Feedforward Network

This paper performs a position-by-position two-layer full connection transformation on $S^{(l)}$, as shown in Formula (12). After repeating the steps for L layers, H is output.

$$\begin{cases} F^{(l)} = \mathrm{ReLU}\left(S^{(l)} W_1^{(l)} + b_1^{(l)} \right) W_2^{(l)} + b_2^{(l)}, \\ X^{(l)} = \mathrm{LayerNorm}\left(S^{(l)} + \mathrm{Dropout}(F^{(l)}) \right). \end{cases} \tag{12}$$

In Formula (12), $W_1^{(l)}$ and $W_2^{(l)}$ represent the weight matrices of the first and second layers, and $b_1^{(l)}$ and $b_2^{(l)}$ represent the corresponding bias terms. In this model, the "sequence" does not refer to the time dimension or text tag sequence, but rather to the feature sequence formed by expanding the fused feature vector Z of a single sample into its dimensionality. The d-dimensional feature vector Z obtained by gated fusion in this paper is reconstructed into a feature sequence of length d (N=d), where each "time step" t corresponds to one dimension of the feature vector. This design enables the Transformer to model higher-order interaction relationships between feature dimensions, rather than traditional temporal dependencies. The positional encoding PE(t) provides the model with relative positional information of feature dimensions, helping to distinguish the semantic roles of different feature channels. The "sequence" in the CRF layer specifically refers to the logical transition relationship between the nine joint label states. Its transition probability is modeled through a learnable matrix A, ensuring that the prediction results conform to the inherent

logical constraint of "satisfaction-turnover intention" (such as avoiding unreasonable jumps such as "high satisfaction-high turnover intention"), rather than continuous observation of samples on the time axis. This two-layer sequence modeling mechanism (feature dimension sequence + label state sequence) effectively solves the joint optimization problem of feature interaction and label coupling in static questionnaire data.

C. CRF MODEL

1) Label State Definition and Emission Score Calculation

This paper applies the CRF model [14,15] to avoid the logical contradiction caused by independent classification by globally modeling the dependency between label sequences. The sequence length is N; the number of joint states is M; the Cartesian product of the three levels of job satisfaction and the three levels of turnover tendency is 9.

The study uses linear mapping to project ht onto an M-dimensional emission vector, as shown in Formula (13).

$$e_t = W_c h_t + b_c \tag{13}$$

In Formula (13), Wc represents the emission weight matrix; bc represents the emission bias vector; et represents the emission score of the state at time t.

To ensure clarity in the state definitions and transition relationships of the CRF layer, this study lists nine joint states of job satisfaction and turnover intention as shown in Table 1. The CRF layer encodes the legitimate combinations of labels through a learnable state transition matrix, penalizing unreasonable jumps (such as jumping directly from "high satisfaction - low turnover intention" to "low satisfaction - high turnover intention"), thus ensuring the logical consistency and rationality of the sequence prediction results.

TABLE 1. Joint Label Definitions for Job Satisfaction and Turnover Intention

Joint state number	Job Satisfaction	Turnover tendency	Tag Description
1	High	Low	High satisfaction - low turnover intention
2	High	Medium	High satisfaction - medium turnover intention
3	High	High	High satisfaction - high turnover intention
4	Medium	Low	Medium satisfaction - low turnover intention
5	Medium	Medium	Medium satisfaction - medium turnover intention
6	Medium	High	Medium satisfaction - high turnover intention
7	Low	Low	Low satisfaction - low turnover intention
8	Low	Medium	Low satisfaction - medium turnover intention
9	Low	High	Low satisfaction - high turnover intention

This study defines "high job satisfaction - high turnover intention" and "low job satisfaction - low turnover intention" as potential conflict combinations, not based on absolute value judgments, but on statistical patterns observed in recent graduates . In the 2512 samples tested, only 3.2% of the cases belonged to the "high job satisfaction - high turnover intention" pattern, and 2.7% belonged to the "low job satisfaction - low turnover intention" pattern. This low-frequency phenomenon is consistent with the theoretical framework; in the early career stages of labor-intensive industries, job satisfaction and turnover intention are typically strongly negatively correlated. The transition matrix A of the CRF learns these statistical patterns through a data-driven approach, rather than by pre-setting rigid rules. During optimization, the transition scores are automatically adjusted by gradient updates, allowing the model to retain rare but reasonable combinations based on the strength of evidence. The use of "potential conflict" in this paper acknowledges the complexity of the situation (such as family factors, career planning, etc.). When external conditions change (such as feature loss), the model automatically reduces the penalty intensity for these combinations; this design balances industry statistical patterns with individual differences. The joint label definitions for job satisfaction and turnover intention are shown in Table 1.

The parallel emission matrix of the entire sequence is expressed as shown in Formula (14).

$$E = [e_1; \ldots; e_N] \tag{14}$$

In Formula (14), E represents the entire emission matrix.

This study uses CRF not to model time-series dependencies, but to structurally model the logical constraints of static joint labels. Although job satisfaction and turnover intention are static variables, they exhibit a strong negative correlation; high satisfaction typically corresponds to low turnover intention, and vice versa. This inherent logical relationship means that the nine joint states are not equally probable: reasonable states (such as "high satisfaction - low turnover intention") should be encouraged, while contradictory states (such as "high satisfaction - high turnover intention") are rare in reality and should be penalized. The transition matrix A of the CRF does not capture temporal dynamics, but rather encodes the semantic compatibility rules between labels, avoiding logical conflicts caused by local predictions through global optimization.

2) Transfer Score and Conditional Probability

In the CRF model of this paper [16,17], the learnable transfer matrix is optimized by gradient update, encoding the legal combination between labels and penalizing unreasonable jumps, representing the score from one state to another.

The calculation of the overall score is shown in Formula (15).

$$s(H, y) = \sum_{t=1}^{N} \left(A_{y_{t-1}, y_t} + e_t[y_t] \right) \tag{15}$$

In Formula (15), s(H,y) represents the overall score; et[yt] represents the emission score of the label yt assigned at time t; y represents the label sequence. Based on the calculation of the overall score, the conditional probability calculation is performed, as shown in Formula (16).

$$p(y \mid H) = \frac{\exp(s(H, y))}{\sum_{\tilde{y} \in Y} \exp(s(H, \tilde{y}))} \tag{16}$$

In Formula (16), Y represents the set of all possible label sequences, and p(y|H) represents the conditional probability.

3) Log-Likelihood Loss

During training, the negative log-likelihood loss is minimized, and the calculation is shown in Formula (17).

$$L_{\text{CRF}} = -\log p(y^* \mid H) = -s(H, y^*) + \log Z(H) \tag{17}$$

In Formula (17), LCRF represents the negative log-likelihood loss. The calculation of log Z (H) in Formula (17) is shown in Formula (18).

$$Z(H) = \sum_{\tilde{y} \in Y} \exp(s(H, \tilde{y})) \tag{18}$$

In Formula (18), Z(H) represents the normalization constant.

4) Viterbi Decoding

In the prediction stage, this paper uses the Viterbi dynamic programming algorithm [18,19] to find the optimal path, and the steps are shown in Formulas (19) to (21).

$$\delta_1(j) = A_{0,j} + e_1[j], \quad \psi_1(j) = 0 \tag{19}$$

$$\begin{aligned} \delta_t(j) &= \max_{1 \le i \le M} \left(\delta_{t-1}(i) + A_{i,j} + e_t[j] \right), \\ \psi_t(j) &= \arg\max_i \left(\delta_{t-1}(i) + A_{i,j} \right) \end{aligned} \tag{20}$$

$$\begin{aligned} y_N^* &= \arg\max_j \delta_N(j), \\ y_t^* &= \psi_{t+1}(y_{t+1}^*), \quad t = N-1, \ldots, 1 \end{aligned} \tag{21}$$

$\delta_t(j)$ represents the optimal cumulative score of the final state j at time t; $\psi_t(j)$ represents the backtracking pointer; y_t^* represents the final output optimal label sequence. The visualization of the optimal path found by the Viterbi dynamic programming algorithm is shown in Figure 1.

In Figure 1, the cumulative score distribution and optimal path decision process calculated by the Viterbi algorithm on a sequence of length T are shown. The grayscale waterfall matrix in Figure 1 corresponds to the cumulative score δ(j,t) for each time step t and state j. The darker the color, the higher the score. The light blue arrows mark the transition decisions from the previous state i to the current state j in an interval manner. The direction and length of the arrows reflect the flow and weight of the transition score A(i,j) in the dynamic programming process. The red broken line and circular mark highlight the optimal state path y* selected at each time step in the final Viterbi decoding, and mark the state index at the node, which intuitively presents how the model balances the emission score and the transition score in the global optimal framework and selects the label sequence with the highest overall score.

D. MODEL FUSION

In this study, the model fusion adopts a hierarchical cascade method, using the Transformer module as a feature encoder, which is responsible for context dependency modeling of the input multi-source data and outputting the deep representation vector of each time step. Process: after completing feature embedding and position encoding, Transformer models the input sequence through a multi-layer encoder and outputs a feature representation sequence containing global dependency information. This sequence is used as the input of the CRF layer to provide sufficient contextual semantic support for subsequent label decoding. The structure of the Transformer-CRF joint prediction model is shown in Figure 2.

In Figure 2, the fused input sequence is fed into the stacked multi-layer Transformer encoder, which captures the global contextual relationship through the multi-head self-attention mechanism and outputs a high-dimensional representation. The CRF decoder constructs the emission score based on H and combines the label transfer matrix

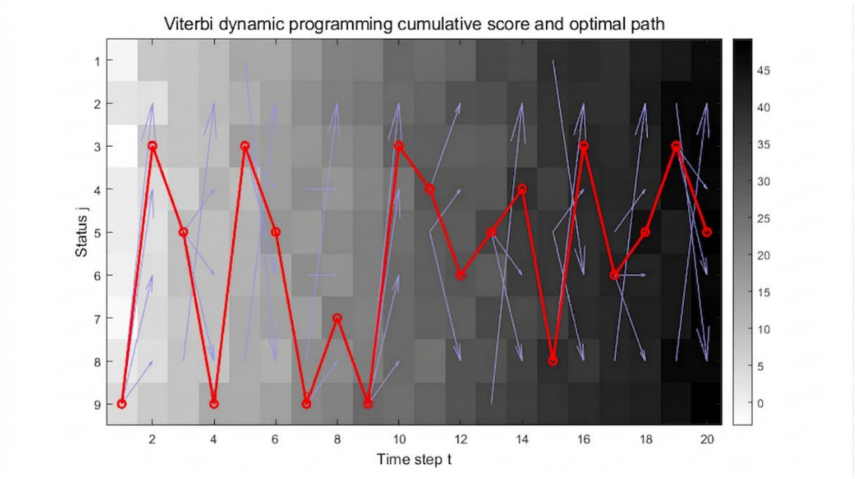

FIGURE 1. Visualization of the optimal path for the Viterbi dynamic programming algorithm

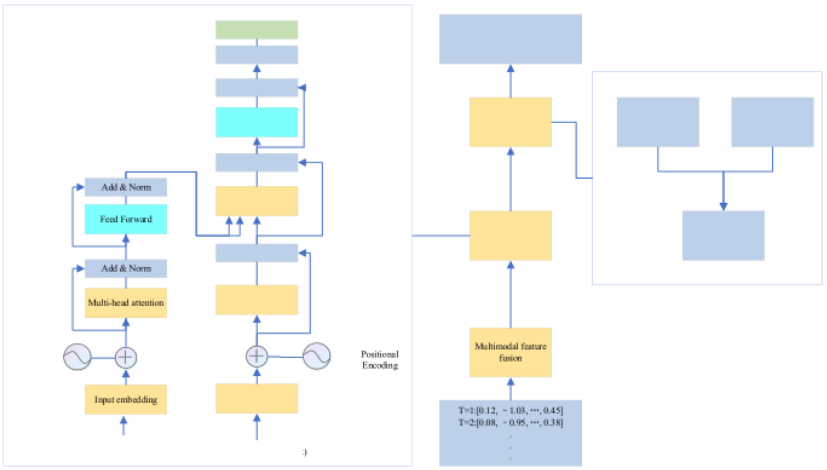

FIGURE 2. Transformer-CRF joint prediction model structure

A, and the Viterbi algorithm is used to infer the optimal label path to achieve joint prediction of job satisfaction and turnover tendency.

E. JOINT TRAINING AND OPTIMIZATION

After completing the model structure design, the experiment adopts an end-to-end joint training strategy, while minimizing the CRF negative log-likelihood and single-task cross-entropy loss, and combines regularization with adaptive learning rate scheduling to ensure that Transformer-CRF is both globally consistent and can achieve stable convergence when jointly predicting two types of labels: job satisfaction and turnover tendency.

1) Loss Function Design

The experimental loss function includes CRF negative log-likelihood and single-task cross-entropy loss. The negative log-likelihood loss of the CRF is shown in Formula (17). The expressions of the single-task cross entropy loss are shown in Formulas (22) and (23).

$$L_{\text{satis}} = -\frac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{3} I[y_t^{\text{satis}} = k]\log(\hat{p}_{t,k}^{\text{satis}}) \tag{22}$$

$$L_{\text{turn}} = -\frac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{3} I[y_t^{\text{turn}} = k]\log(\hat{p}_{t,k}^{\text{turn}}) \tag{23}$$

I represents the indicator function; y_t^{satis} and y_t^{turn} represent the true satisfaction and turnover tendency labels, respectively; $\hat{p}_{t,k}^{\text{satis}}$ and $\hat{p}_{t,k}^{\text{turn}}$ are the Softmax probabilities corresponding to the k classes. The total joint loss is shown in Formula (24).

$$L = \alpha L_{\text{CRF}} + \beta_1 L_{\text{satis}} + \beta_2 L_{\text{CE}} + \frac{\lambda}{2}\|\Theta\|_2^2 \tag{24}$$

α, β_1, and β_2 represent the weight hyperparameters of the CRF and the two single-task cross entropy losses, respectively. λ represents the L2 regularization parameter, and Θ represents the set of all learnable parameters.

2) Joint Optimization

This paper uses the AdamW optimizer for training optimization [20-22], and the update rule is shown in Formula (25).

$$\Theta_t = \Theta_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon_1} \tag{25}$$

In Formula (25), η_t represents the learning rate; ϵ_1 represents the smoothing constant; $\hat{m}_t$ and $\hat{v}_t$ represent the first-order and second-order moment estimates, respectively.

To speed up the training, the linear warm-up + cosine annealing strategy is used to adjust the learning rate, as shown in Formula (26).

$$\eta_t = \begin{cases} \dfrac{t}{T_{\text{warmup}}}\eta_{\max}, & t \le T_{\text{warmup}}, \\ \eta_{\min} + \dfrac{1}{2}(\eta_{\max} - \eta_{\min}) & \\ \quad \times \left(1 + \cos\left(\dfrac{\pi(t - T_{\text{warmup}})}{T_{\text{total}} - T_{\text{warmup}}}\right)\right) & t > T_{\text{warmup}}. \end{cases} \tag{26}$$

In Formula (26), T_{warmup} represents the number of warm-up steps; T_{total} represents the total number of training steps; $\eta_{\max}$ and $\eta_{\min}$ represent the initial maximum learning rate and minimum learning rate, respectively.

To prevent gradient explosion, the experiment performs global norm clipping on all parameter gradients and applies Dropout and LayerNorm after the Transformer and CRF emission layers to enhance the robustness of the model. The experimental training hyperparameters are shown in Table 2.

TABLE 2. Experimental training hyperparameters

Parameters	Value
Learning rate	0.0003
Batch size	32
Maximum number of training epochs	20
Number of Transformer layers	6
Number of attention heads	8
Hidden dimensions	256
Dropout probability	0.1
CRF label set size	9
Optimizer	AdamW
Weight decay	0.01

IV. JOINT PREDICTION EXPERIMENT OF JOB SATISFACTION AND TURNOVER TENDENCY OF COLLEGE GRADUATES

A. EXPERIMENTAL DATA

The experimental data of this paper comes from the online employment satisfaction and turnover tendency survey conducted by Nanjing University's 2023 graduates, and a total of 2,512 valid questionnaires are collected. The data types include category feature data, numerical feature data, and text feature data. In the category feature data, there are 6 fields, including gender (male/female), major category (science/literature/business/arts), education level (undergraduate/master/doctor), and position type (technology/management/sales/administration), Company type (state-owned enterprise/private enterprise/foreign-invested enterprise/public institution) and work location (eastern coastal/central/western region). In the numerical feature data, there are 7 scale scores, each with a score of 1-5. Other indicators include age at work (months), monthly salary level (thousands of yuan), and average weekly overtime hours (hours). In the text feature data, there are 5,024 short texts, including graduates' self-reports on their onboarding experience, team atmosphere, and career development expectations. The text feature data design provided each graduate with two independent short texts: the first required a description of "the most profound work experience within the first three months of employment" (average length 68 words), and the second required an explanation of "the most critical factor affecting job stability" (average length 72 words). Therefore, 2512 valid questionnaires generated a total of 5024 short texts, covering graduates' self-evaluations of their onboarding experience, team atmosphere, and career development expectations.

The seven scales used in this paper are: (1) the short version of the Minnesota Job Satisfaction Scale (MSQ), which measures internal/external satisfaction (α = 0.87); (2) the Mobley Turnover Intention Scale, which includes three items to measure the willingness to leave a job voluntarily (α = 0.82); (3) the Luthans Psychological Capital Scale (PCQ-24), which assesses self-efficacy/hope/resilience/optimism (α = 0.89); (4) the Meyer Affective Commitment Scale, which measures organizational attachment strength (α = 0.85); (5) the Occupational Identity Scale, which assesses professional identity (α = 0.79); (6) the Perceived Work Stress Scale (PSS-10), which measures role conflict and workload (α = 0.83); and (7) the Perceived Pay Fairness

Scale, which includes three dimensions: distributional/procedural/interactional fairness (α = 0.86). All scales used a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree), with scoring rules following industry standards: 1-2 points were defined as "low", 3 points as "moderate", and 4-5 points as "high". Pre-testing of the scales was conducted with 200 non-sample graduates, and confirmatory factor analysis showed good fit indices for each scale (CFI > 0.93, RMSEA < 0.07).

This study strictly adhered to research norms, and all participants signed electronic informed consent forms before completing the questionnaires. The sampling framework covered 6,823 graduating students from all 12 faculties of Nanjing University in 2023, employing a stratified random sampling strategy: first stratified by academic discipline (science/literature/business/arts), and then allocating the sample size proportionally by department size. The questionnaires were distributed in three rounds (7-day intervals) through the official platform of the university's Career Guidance Center, ultimately collecting 3,148 responses. After rigorous quality control (completion rate <85%, logical contradictions ≥3,

open text <20 characters or duplication rate >30%), 636 invalid questionnaires were eliminated, resulting in an effective response rate of 36.8%. A privacy protection mechanism was specifically implemented for text data collection, automatically masking sensitive information such as ID numbers, mobile phone numbers, and specific salary figures. Two research assistants cross-verified the anonymization effect. The final text database was stored on the university's internal server using hash encryption, retaining only the semantic feature vectors required for analysis. The original text was permanently deleted within 72 hours after feature extraction. The sample characteristic distribution was not significantly different from the population distribution by the chi-square test ($p > 0.05$), ensuring the representativeness of the data.

B. DATA PREPROCESSING

The experiment fills and removes missing values and outliers in the original questionnaire. To strictly avoid data leakage, all missing value imputation was performed independently within a 10-fold cross-validation framework. In each fold experiment, the similarity matrix of the KNN (K-Nearest Neighbor) interpolator was calculated using only the current training set (90% of the samples), and the missing values in the test set (10% of the samples) were imputed using the statistical features of the training set. For missing numerical features in the training set, the KNN interpolator (nearest neighbor count k = 5) was used to find the 5 most similar complete samples in the same batch based on Euclidean distance for mean imputation. For missing values in the test set, linear interpolation was performed using the standardized mean and standard deviation of the corresponding features from the training set.

This paper uses the Jieba Chinese word segmentation tool to segment 5,024 short texts and uses the stop word list to remove high-frequency non-informative words. Based on TF-IDF, 2,000-dimensional sparse feature vectors are extracted, and PCA is used to retain the principal components with a cumulative variance contribution of 95%, reducing the feature dimension to 200.All dimensionality reduction parameters (PCA principal component vectors) are calculated only on the training set, and the test set features are transformed through the same projection matrix to ensure the unbiasedness of the evaluation results.

C. EXPERIMENTAL PLAN

To comprehensively evaluate the performance of the model in the joint prediction task of job satisfaction and turnover tendency of college graduates, this study uses ten-fold cross validation to divide the data into training and test sets. To ensure fairness in the comparison, all models were trained on the same multi-source fusion features. First, a standardized feature matrix Z of uniform dimension was generated. For models originally designed for plain text (such as RoBERTa-CRF), their input layer structure was modified by removing the original word embedding layer. The fusion features Z were then mapped to the model's desired hidden dimension (256-dimensional) via linear projection, while retaining the pre-trained Transformer parameters for feature enhancement. Non-sequence models such as XGBoost directly received Z as the input vector. BiLSTM variants also received a serialized representation of Z (expanding the 256-dimensional features into a sequence of length 256). All models underwent 10-fold cross-validation under the same training/test split, and hyperparameter optimization was strictly limited to the training set.All comparison models, including XGBoost based on gradient boosting, single-task models using only Transformer encoders, BiLSTM-CRF, RoBERTa-CRF (Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach-Conditional Random Field), BiLSTM-CRF-CNN, and Transformer-CRF in this paper, are trained under the same data partitioning and preprocessing process to ensure the comparability of results. The hyperparameter search uses a combination of grid search and Bayesian optimization to determine the optimal learning rate, hidden layer dimension, regularization coefficient, and Dropout rate for each model. The benchmark models selected in this paper include XGBoost, Transformer, BiLSTM-CRF, RoBERTa-CRF, and BiLSTM-CRF-CNN, primarily for two reasons. XGBoost is a commonly used high-performance non-sequence HR prediction method that reflects the advantages of traditional machine learning in feature processing. Transformer, BiLSTM-CRF, and RoBERTa-CRF are typical sequence modeling methods, widely used in text and behavioral sequence analysis, and can capture the temporal dependence features of employee satisfaction and turnover intention. Therefore, these models represent both the mainstream methods of current non-joint prediction and typical deep sequence models, providing a reasonable and representative reference for the joint prediction method proposed in this paper.

D. EVALUATION INDICATORS

The complete match rate is shown in Formula (27).

$$\mathrm{EMR} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i = y_i) \tag{27}$$

In Formula (27), $\hat{y}_i$ represents the predicted joint label, and y_i represents the actual joint label. The label conflict rate is shown in Formula (28).

$$\mathrm{LCR} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i \in C) \tag{28}$$

In Formula (28), C represents the predefined "conflicting" joint label set (unreasonable combinations such as "high satisfaction-high turnover tendency")

The joint label consistency rate is shown in Formula (29).

$$\mathrm{JCR} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i \in L) \tag{29}$$

In Formula (29), L represents the set of all legal (non-conflicting) joint labels.

The inter-label dependency MI is shown in Formula (30).

$$\mathrm{MI}(S_a, TL) = \sum_{k=1}^{K}\sum_{l=1}^{L} p(s_{a,k}, tl_l)\log\left(\frac{p(s_{a,k}, tl_l)}{p(s_{a,k})p(tl_l)}\right) \tag{30}$$

In Formula (30), Sa represents the satisfaction label, and TL represents the turnover tendency label. Cohen's Kappa coefficient is shown in Formulas (31) and (32).

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{31}$$

$$p_o = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i = y_i), \quad p_c = \sum_{u=1}^{M}\left(\frac{n_u}{N}\right)\left(\frac{m_u}{N}\right) \tag{32}$$

po represents the observed consistency rate; pc represents the random consistency rate; nu represents the number of samples in the test set whose true label is equal to the u-th class; mu represents the number of samples in the test set whose predicted label is equal to the u-th class.

V. JOINT PREDICTION RESULTS OF JOB SATISFACTION AND TURNOVER TENDENCY

A. CONFUSION MATRIX

All experimental results were generated based on a rigorous 10-fold cross-validation framework. The original dataset (2512 samples) was divided into 10 subsets using stratified sampling: 6 folds (252 samples) and 4 folds (251 samples) (total 2512). In each fold experiment, 9 folds (2260±1 samples) were used as the training set for model training and hyperparameter tuning (internally further divided into 8:1 ratios for early stopping), and the remaining 1 fold was used as an independent test set for evaluation. All metrics in the experimental report are macro-means of the 10-fold test results. Note: The confusion matrix is visualized using the 7th fold (252 samples) as an example, as its class distribution is closest to the overall dataset (chi-square test p = 0.92), but quantitative analysis is still based on the 10-fold average results. This design ensures statistical rigor while enhancing the interpretability of the results through visualization of typical folds. On the test set (252 samples), the prediction results of the model are intuitively presented through the confusion matrix to show the classification performance of each joint label. The confusion matrix is shown in Figure 3. In Figure 3, the categories include high satisfaction-low tendency (45), high satisfaction-medium tendency (30), high satisfaction-high tendency (15), medium satisfaction-low tendency (50), medium satisfaction-high tendency (20), medium satisfaction-medium tendency (41), low satisfaction-low tendency (30), low satisfaction-medium tendency (15), and low satisfaction-high tendency (6). The horizontal direction represents the predicted label, and the vertical direction represents the actual label.

In the 252 samples in the test set of Figure 3, the model performs well for most common categories. There are 45 actual samples in the "high satisfaction-low tendency" category, of which 35 are correctly predicted, with an accuracy of 77.8%; 36 out of 50 samples of "medium satisfaction-low tendency" are correctly predicted, with an accuracy of 72.0%; 30 out of 41 samples of "medium satisfaction-medium tendency" are correctly predicted, with an accuracy of 73.2%. For categories with fewer samples, such as "high satisfaction-high tendency" and "ow satisfaction-high tendency", the number of correct predictions is 11 and 4, respectively, corresponding to accuracy rates of 73.3% and 66.7%. Overall, the average accuracy rate of categories with larger sample sizes (≥ 30) is about 72.6%, while the average accuracy rate of categories with smaller sample sizes (≤ 15) is only about 68.9%, which shows a certain decline in cold category performance.

The sample size of each joint label is unevenly distributed. The number of samples of the category with mixed tendency of medium satisfaction and high satisfaction is relatively small. There are only 15 samples of "high satisfaction-high tendency". Due to the small number of corresponding gradient updates during training, the model is difficult to fully learn the discriminant features of this category, resulting in 2 cases being misjudged as "medium satisfaction-high tendency" and 1 case being judged as "high satisfaction-medium tendency". There are only 6 samples of "low satisfaction-high tendency"; the remaining 1 case is misjudged as "medium satisfaction-high tendency", and 1 case is "low satisfaction-medium tendency". This inaccurate estimation of intra-category variance caused by the scarcity of samples makes the model fluctuate more on the boundary features.

B. JOINT PREDICTION PERFORMANCE

The comparison results of the joint prediction performance of each model on the test set are shown in Figure 4.

In Figure 4, in terms of complete match rate and joint prediction precision, XGBoost only reaches 82.3% and 85.1%, respectively, while Transformer-CRF increases to 96.6% and 96.1%, respectively, an increase of 14.3% and 11.0% compared with XGBoost. BiLSTM-CRF-CNN achieves a complete match rate of 93.4% and a precision of 94.0%, and RoBERTa-CRF achieves a complete match rate and precision of 91.5% and 93.2%, respectively. It can be seen that the Transformer-CRF prediction output has the best proportion of completely correct matching labels and the best prediction precision.

In terms of joint recall and joint prediction F1 indicators, XGBoost is 83.8% and 84.4%, respectively, while Transformer-CRF is 95.2% and 95.6%, an increase of about 11.4% and 11.2%. Transformer has a recall of 88.1% and F1 of 88.8%; BiLSTM-CRF-CNN is 92.8% and 93.4%; RoBERTa-CRF is 91.9% and 92.5%.

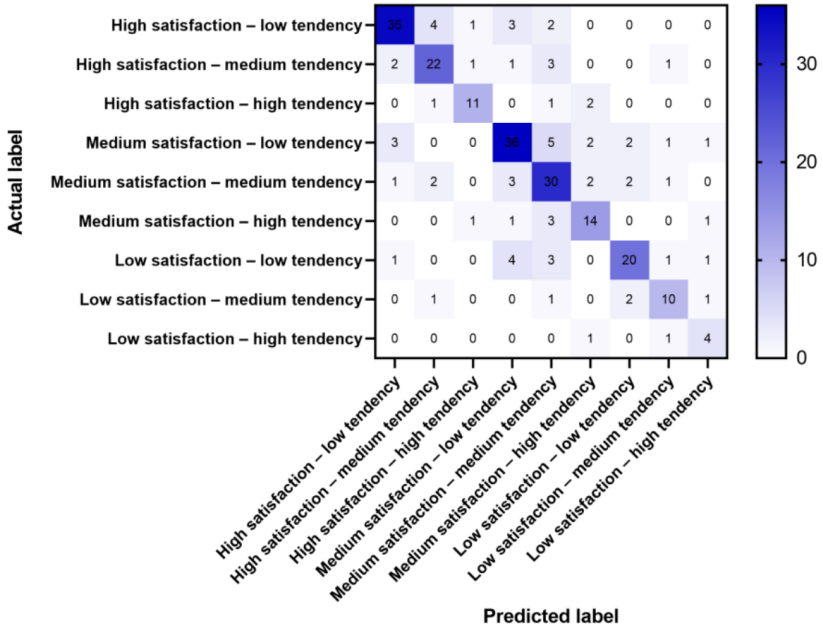

FIGURE 3. Confusion matrix

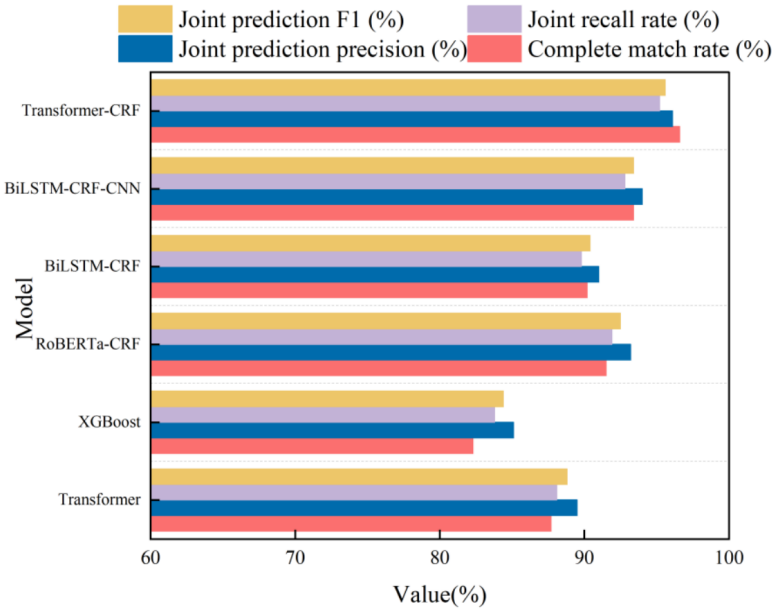

FIGURE 4. Joint prediction performance

C. JOINT LABEL CONSISTENCY

To evaluate the consistency of joint labels, the experiment counts the label conflict rate and the joint label consistency rate, and calculates the inter-label dependency MI and Kappa coefficient to measure the correction effect of annotation alignment and random consistency. The results of joint label consistency are shown in Table 3.

TABLE 3. Joint label consistency results

Model	Label conflict rate (%)	Joint label consistency rate (%)	Inter-label dependency index MI	Kappa coefficient
XGBoost	12.3	87.7	0.3	0.66
Transformer	8.2	91.8	0.4	0.74
BiLSTM-CRF	6.5	93.5	0.4	0.78
RoBERTa-CRF	4.1	95.9	0.5	0.83
BiLSTM-CRF-CNN	3.3	96.7	0.5	0.86
Transformer-CRF	1.8	98.2	0.6	0.89

In the comparison of label conflict rate and joint label consistency rate in Table 3, XGBoost has the highest conflict rate of 12.3%, and the corresponding consistency rate is only 87.7%, indicating that more than 10% of the predictions fall into unreasonable combinations outside the predefined “high satisfaction-high tendency” or “low satisfaction-low tendency”. Transformer has a conflict rate of 8.2% and a consistency rate of 91.8%, while other models BiLSTM-CRF (6.5%, 93.5%), RoBERTa-CRF (4.1%, 95.9%), and BiLSTM-CRF-CNN (3.3%, 96.7%) are all worse. Transformer-CRF reduces the conflict rate to 1.8% and increases the consistency rate to 98.2%, which shows that almost all predictions meet the logical rationality between “satisfaction-tendency”.

In terms of the inter-label dependency MI and Kappa coefficient, XGBoost’s MI is only 0.30, and Kappa is 0.66, indicating that the dependency between its predicted labels and true labels is weak, and the consistency with random predictions is low. Transformer has an MI of 0.40 and a Kappa of 0.74, while BiLSTM-CRF and RoBERTa-CRF achieve MI 0.40/0.50 and Kappa 0.78/0.83, respectively; BiLSTM-CRF-CNN achieves MI 0.50 and Kappa 0.86. The Transformer-CRF model achieves MI 0.60 and Kappa 0.89, which shows that the model not only better captures the joint distribution characteristics of the “satisfaction” and “turnover tendency” labels in a statistical sense, but also makes significant progress in correcting random consistency.

D. COMPARISON OF SINGLE-TASK CLASSIFICATION PERFORMANCE

At the single-task classification level, to further compare the performance of the models on their respective subtasks, the accuracy, precision, recall, and F1 of the two indicators of job satisfaction and turnover tendency are evaluated, respectively, and the results are shown in Figure 5.

Figure 5(a). Job satisfaction classification performance

Figure 5(b). Turnover tendency classification performance

In the job satisfaction subtask in Figure 5(a), XGBoost’s accuracy is only 81.7%; precision is 83.2%; recall is 80.5%; F1 value is 81.8%. The Transformer-CRF model achieves 94.5% accuracy, 95.0% precision, 94.0% recall, and 94.5% F1, which are 12.8, 11.8, 13.5, and 12.7 percentage points higher than XGBoost, respectively. RoBERTa-CRF (91.3%, 92.0%, 90.6%, 91.3%) and BiLSTM-CRF-CNN (92.0%, 92.7%, 91.2%, 92.0%) also exceed 90%, indicating that the Transformer-CRF model has strong complex features and sequence dependency modeling capabilities in job satisfaction prediction, and the overall performance has been greatly improved. In the turnover tendency subtask in Figure 5(b), the baseline XGBoost index is slightly lower than the satisfaction prediction, with an accuracy of 79.3%, a precision of 80.1%, a recall of 78.4%, and an F1 of 79.2%. Transformer-CRF model also achieves significant improvements on this task, reaching 93.2% accuracy, 93.8% precision, 92.6% recall, and 93.2% F1, which are 13.9, 13.7, 14.2, and 14.0 percentage points higher than XGBoost, respectively. RoBERTa-CRF (89.5%, 90.1%, 88.8%, 89.4%) and BiLSTM-CRF-CNN (90.8%, 91.4%, 90.2%, 90.8%) achieve good results, but are still slightly worse than the performance of satisfaction prediction, reflecting that the label signal of turnover tendency is more dispersed and more dependent on the deep fusion of multi-source heterogeneous features.

E. ATTENTION WEIGHT DISTRIBUTION

To show the difference in the degree of attention paid to each feature in the joint prediction task, the key influencing factors are analyzed. The distribution of attention weights of the Transformer multi-head attention mechanism on different feature dimensions is shown in Figure 6.

Figure 6(a). Attention heat map

Figure 6(b). Average attention bar graph

The model’s attention heat map (Figure 6a) shows significant differences in how attention heads focus on features, confirming that different heads capture distinct feature subspaces within the multi-source data. For instance, the third head heavily weights "psychological capital-self-efficacy" and "-resilience" (0.167, 0.168), virtually ignoring "professional category" (0.004). Conversely, the sixth head strongly focuses on "organizational commitment-emotional commitment" (0.185). The first head, meanwhile, prioritizes the "text semantic vector" (0.128).

The average attention bar chart (Figure 6b) provides a smoother representation of overall feature importance. The most discriminative inputs for the joint label prediction are "psychological capital-resilience" (highest average weight: 0.109) and "average weekly overtime hours" (0.097). The "text semantic vector" is also highly important (0.084). Features with the lowest influence include "psychological capital-optimism" (0.043) and "age of employment" (0.048). This suggests that the model primarily uses the resilience index and actual working hours information from the quantitative scale as the most predictive factors for the joint satisfaction and turnover label.

F. TRAINING EFFICIENCY AND RESOURCE CONSUMPTION

To evaluate the performance of each model in terms of training efficiency and system resource usage, this paper counts the training time, average response time, CPU (Central Processing Unit) utilization, and throughput of six types of models under the same hardware environment. The training efficiency and resource consumption are shown in Table 4.

TABLE 4. Training efficiency and resource consumption

Model	Training time (s)	Average response time (ms)	CPU utilization (%)	Throughput (samples/s)
XGBoost	12.6	8.3	31.4	121.3
Transformer	108.4	22.7	68.9	47.5
BiLSTM-CRF	134.7	26.1	72.6	39.8
RoBERTa-CRF	165.3	30.4	75.8	33.6
BiLSTM-CRF-CNN	142.1	28.6	73.4	36.1
Transformer-CRF	116.9	23.9	70.2	44.8

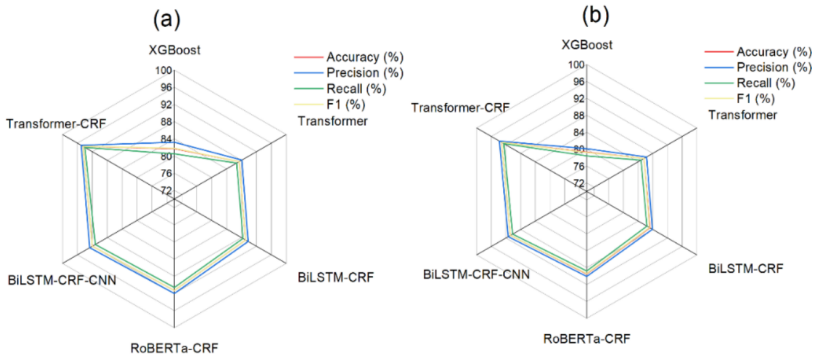

FIGURE 5. Single-task classification performance

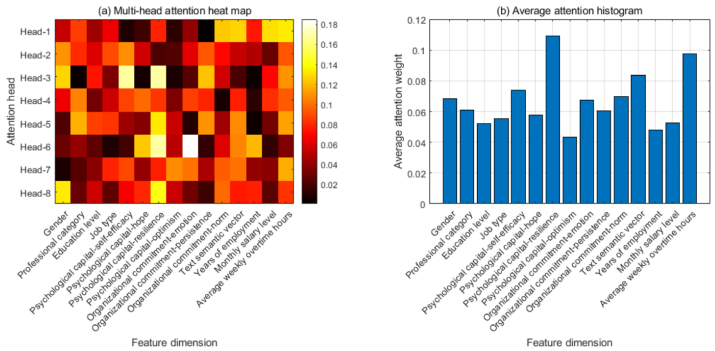

FIGURE 6. Attention weight distribution

In terms of efficiency (Table 4), the non-neural XGBoost model is the clear winner, requiring only 12.6 seconds for training and an average response time of 8.3 ms, with the highest throughput of 121.3 samples/s due to its minimal computational overhead. Deep learning models require more resources because of their sequence dependency and context modeling structures. The resource-intensive RoBERTa-CRF has the worst efficiency, with the longest training time (165.3 seconds) and highest response latency (30.4 ms), largely due to loading large pre-trained parameters. The proposed Transformer-CRF model strikes a balance, offering better efficiency among the deep learning approaches: its training time (116.9 seconds) and response time (23.9 ms) are competitive, and its throughput (44.8 samples/s) is superior to both BiLSTM-CRF and BiLSTM-CRF-CNN. This is because Transformer-CRF relies on the attention mechanism rather than slower recurrent structures, demonstrating a better controlled resource consumption while maintaining strong expressive power.

G. ROBUSTNESS TEST

To verify the stability and robustness of the Transformer-CRF model under different disturbance conditions, this paper designs robustness test experiments from three typical interference scenarios: Gaussian noise disturbance in input data, changes in the proportion of unbalanced label distribution, and feature missing scenarios. The robustness test results are shown in Figure 7.

Figure 7(a). Robust performance under Gaussian noise perturbation

Figure 7(b). Robust performance under changes in the proportion of unbalanced label distribution

Figure 7(c). Robust performance in the case of feature loss

The Transformer-CRF model's robustness was evaluated under three stress tests.

First, the Gaussian noise perturbation test (Figure 7a) revealed the model's sensitivity to input data corruption. A low standard deviation (σ = 0.01) resulted in a near-perfect 95.9% complete match rate and 97.9% joint consistency. However, increasing σ to 0.2 significantly degraded performance, dropping the match rate to 85.3% and increasing the label conflict rate to 6.0%. This demonstrates that noise interferes with the model's ability to accurately capture feature relationships, particularly long-distance dependencies. Second, the label distribution imbalance test (Figure 7b) showed that the model's predictive power weakens as minority class sparsity increases. While a 33% proportion of high turnover tendency samples yielded a 96.6% match rate, reducing the proportion to 13% caused the match rate to plummet to 88.5% and the consistency rate to 92.1%.

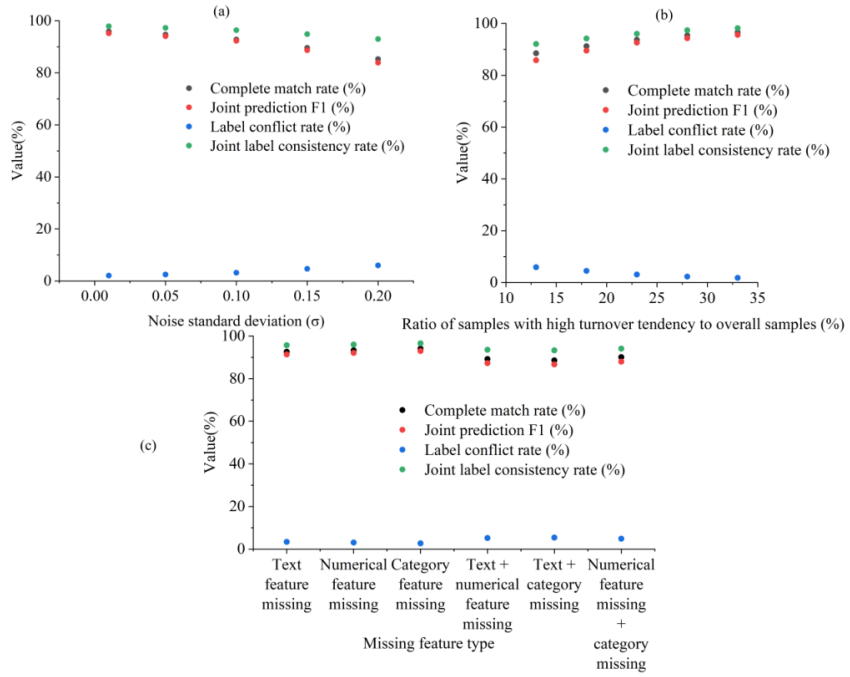

FIGURE 7. Robustness test results

This indicates that sparse minority class samples lead to increased label inconsistency and misclassification, as the model's decision boundaries become fuzzy and minority samples are incorrectly absorbed into the majority class.

Third, the feature missing test (Figure 7c) demonstrated that text features are the most critical, playing a core role in characterizing subjective experience. Missing text features alone dropped the complete match rate to 92.6% and the F1 score to 91.4%, with the consistency rate at 95.7%. Missing numerical features (like psychological capital) and categorical features had comparatively smaller individual impacts (F1 of 92.1% and 93.0% respectively). However, combined feature loss, such as missing text + category (F1 of 86.7%) or text + numerical (F1 of 87.3%), resulted in the most significant degradation.

H. GENERALIZATION TEST

To further verify the generalization ability of the Transformer-CRF model, this paper tests graduate data from different universities and different years. The generalization test results are shown in Figure 8.

Figure 8 (a). Test results of different universities

Figure 8 (b). Test results of graduate data from different years

The Transformer-CRF model exhibits strong generalization ability but shows performance variations across different university types and time periods. In cross-university testing (Figure 8a), the model achieved a 94.3% complete match rate and 93.5% joint F1 at high-level institutions like Zhejiang University, closely matching the training data (Nanjing University). However, performance decreased for colleges with distinct language styles, such as Shenzhen Polytechnic (85.9% match rate, 5.3% label conflict rate), indicating that individual differences in student expression and employment concepts affect generalization. Temporally (Figure 8b), the model shows excellent stability, with the complete match rate consistently rising from 93.8% (2018) to 96.1% (2022) as the data collection year approached the 2023 training set. This trend, sustained by a 95.0% match rate in 2024 data, confirms the model's scalability; variations in socioeconomic environment, job market expectations, and evolving language habits across different periods and institutions are the primary factors affecting feature consistency and structural prediction accuracy.

I. COMPARISON OF MODEL OUTPUT AND REAL MANAGEMENT DECISION EFFECTS

To further evaluate the gain of model output on college employment guidance decisions, this experiment designs a comparative experiment, comparing Transformer-CRF model guidance with experience decisions, mainly to explore the gain of complete match rate improvement on college employment guidance. The experimental steps are as follows:

(1) Data of Nanjing University's 2023 graduates is selected, including the prediction results of students' job satisfaction and turnover tendency. The job satisfaction and turnover tendency data predicted by the model are compared with the traditional experience decision (based on the judgment of teachers and employment guidance departments).

(2) Based on the prediction results (model output) and empirical decision-making (empirical observation and evaluation), the career development and employment stability

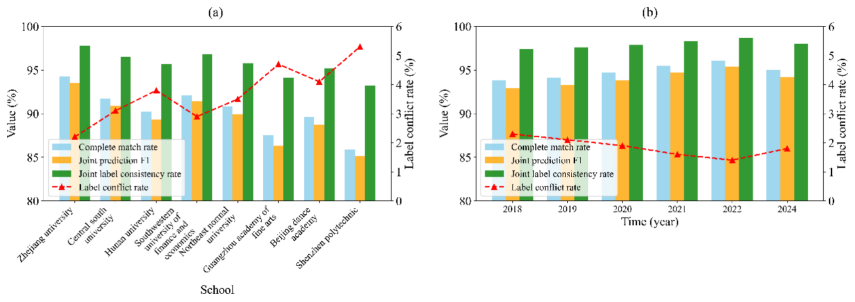

FIGURE 8. Generalization test results

intervention strategies of graduates are guided. Intervention measures include regular career counseling, emotional support, salary negotiation skills improvement, etc.

(3) Through 6 months of tracking data, indicators such as retention rate, satisfaction change, and intervention cost are evaluated, and the effects of the two decision-making methods are compared. The effect of improving the complete match rate on the gains of college employment guidance is shown in Table 5.

TABLE 5. The effect of improving the complete match rate on the gains of college employment guidance

Indicators	Prediction group (model guidance)	Control group (experience decision)	Gain
Complete match rate (%)	96.6	65.0	31.6
Retention rate after 6 months (%)	92.3	85.1	7.2
Average satisfaction improvement (points)	1.6	0.8	0.8
Unit intervention cost (¥/person)	1200	1600	25%

Table 5 clearly shows that the model-guided prediction group significantly outperformed the experience-based control group across all indicators. The model achieved a complete match rate of 96.6%, a remarkable 31.6% increase over the control group's 65.0%, confirming its superior accuracy in predicting job satisfaction and turnover tendency. Furthermore, the prediction group's 6-month retention rate was 92.3% (7.2% higher than the control group's 85.1%), demonstrating that precise prediction and targeted interventions effectively improve employee retention. The model also improved employee experience, with an average satisfaction score of 1.6, compared to only 0.8 in the control group. Crucially, the unit intervention cost for the prediction group was only 1200 yuan per person, a 25% reduction from the 1600 yuan per person in the control group, demonstrating that a precise model-driven approach can significantly reduce costs while improving effectiveness.

VI. CONCLUSION

This paper adopts a joint prediction model of job satisfaction and turnover tendency of college graduates based on Transformer-CRF, specifically tailored for application within the industry where labor dynamics and retention challenges necessitate a highly accurate, coupled analysis of employee sentiment and exit risk. The global context representation is extracted by using a multi-layer self-attention Transformer encoder. The CRF decoder is connected to the top layer of the encoder. The nine-state label sequence of "satisfaction-tendency" is jointly optimized by the emission score and the transfer matrix, and the training is performed by end-to-end minimization of the CRF log-likelihood and auxiliary cross-entropy loss. The experimental results show that the model has a joint label complete match rate of 96.6% and a label conflict rate of only 1.8%, which is significantly better than baseline methods such as XGBoost, BiLSTM-CRF, and RoBERTa-CRF. The model performs stably in robustness testing, unbalanced label distribution, and multi-school and cross-year data verification. Despite the promising results, this paper still has shortcomings such as a single feature source, and its regional and cultural applicability, particularly within diverse operational contexts like the global ndustry, remains to be fully verified, coupled with the challenge of high hyperparameter optimization cost. Future work should therefore focus on incorporating richer data, such as behavioral logs and social network data, conducting rigorous cross-cultural localization research to ensure its relevance for manufacturing hubs across different continents, and exploring neural architecture search (NAS) to improve model automation and efficiency in managing the specialized human re-source challenges of the sector.

AUTHOR CONTRIBUTIONS

Conceptualization – Fang Rao, Wei Cao, Jianxue Huang and Yi Zhou; methodology – Fang Rao, Wei Cao, Jianxue Huang and Yi Zhou; investigation – Fang Rao, Wei Cao, Jianxue Huang and Yi Zhou; writing-original draft preparation – Fang Rao, Wei Cao, Jianxue Huang and Yi Zhou. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received was supported by,

(1) Jiangxi Province Education Science "14th Five-Year Plan" project "Psychological capital on the early career stability of college graduates" (23YB293);

(2) Nanchang Normal University Doctoral Research Start-up Fund project "Psychological capital on the employment stability of college graduates" (NSBSJJ2023001).

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

All eligible participants signed an informed consent form. To ensure the protection of privacy, all data collected were processed under strict protocols for anonymity.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] W. Wang, Y. Lai, J. Deng, J. Shi, and L. You, "Research on High-Quality Employment of College Students from the Perspective of Self-Marketing: A Case Study of Local Institutions in Guangdong," Modern Economy, vol. 15, no. 10, pp. 1000-1025, 2024.

[2] C. Oliveira-Silva, A. Soares-Semedo, and B. Lopez-Bermudez, "Unemployed foreign graduates: job prospects and options–a case analysis in Portugal," Journal of Entrepreneurship and Public Policy, vol. 13, no. 1, pp. 37-57, 2024, doi: 10.1108/JEPP-02-2023-0020.

[3] S. Y. Sari, Z. Yenni, and M. H. Aima, "Determinants of Turnover Intention: Job Satisfaction, Employee Retention, Work-Family Conflict and Organisational Commitment," International Review of Management and Marketing, vol. 14, no. 6, pp. 26-36, 2024, doi: 10.32479/irmm.16979.

[4] A. Gupta, A. Chadha, V. Tiwari, A. Varma, and V. Pereira, "Sustainable training practices: predicting job satisfaction and employee behavior using machine learning techniques," Asian Business & Management, vol. 1, no. 1, pp. 1-24, 2023, doi: 10.1057/s41291-023-00234-5.

[5] O. Espinoza, L. Gonzalez, C. Miranda, L. Sandoval, B. Corradi, N. McGinn, et al., "Job satisfaction among university graduates in Chile," Higher Education, Skills and Work-Based Learning, vol. 14, no. 4, pp. 865-883, 2024, doi: 10.1108/HESWBL-10-2023-0286.

[6] M. Liu, B. Yang, and Y. Song, "Research on Predicting the Turnover of Graduates Using an Enhanced Random Forest Model," Behavioral Sciences, vol. 14, no. 7, pp. 562-574, 2024, doi: 10.3390/bs14070562.

[7] D. Ma, M. Shu, and H. Zhang, "Feature selection optimization for employee retention prediction: A machine learning approach for human resource management," 2025; 10(1):1-11, doi: 10.20944/preprints202504.1549.v1.

[8] W. Li, "A transformer-based deep learning framework to predict employee attrition," PeerJ Computer Science, vol. 9, no. 1, pp. 1-12, 2023, doi: 10.7717/peerj-cs.1570.

[9] B. Town and V. Padmavathy, "Predictive Analysis of Employee Turnover in IT using a Hybrid CRF-BiLSTM and CNN Model," Proceedings of the 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA); 15-17 November 2023; Theni, India. Piscataway, NJ, USA: IEEE; 2023. Volume 7(1), pp. 914-919.

[10] Y. Zhu, "A knowledge graph and BiLSTM-CRF-enabled intelligent adaptive learning model and its potential application," Alexandria Engineering Journal, vol. 91, no. 1, pp. 305-320, 2024, doi: 10.1016/j.aej.2024.02.011.

[11] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," ACM computing surveys, vol. 56, no. 9, pp. 1-36, 2024, doi: 10.1145/3649447.

[12] V. Bhogade and B. Nithya, "Time series forecasting using transformer neural network," International Journal of Computers and Applications, vol. 46, no. 10, pp. 880-888, 2024, doi: 10.1080/1206212X.2024.2396321.

[13] D. K. Kim and K. Kim, "A convolutional transformer model for multivariate time series prediction," IEEE Access, vol. 10, no. 1, pp. 101319-101329, 2022, doi: 10.1109/ACCESS.2022.3203416.

[14] J. S. Park, S. Fadnavis, and E. Garyfallidis, "Multi-scale V-net architecture with deep feature CRF layers for brain extraction," Communications Medicine, vol. 4, no. 1, pp. 29-41, 2024, doi: 10.1038/s43856-024-00452-8.

[15] M. Yin, P. Zhou, T. Xu, and J. Jiang, "Extracting actors and use cases from requirements text with BiLSTM-CRF," Journal of Intelligent & Fuzzy Systems, vol. 44, no. 3, pp. 4285-4299, 2023, doi: 10.3233/JIFS-221094.

[16] Z. Sun and X. Li, "Named entity recognition model based on feature fusion," Information, vol. 14, no. 2, pp. 133-145, 2023, doi: 10.3390/info14020133.

[17] Y. Yan, F. Liu, X. Zhuang, and J. Ju, "An R-transformer_BiLSTM model based on attention for multi-label text classification," Neural Processing Letters, vol. 55, no. 2, pp. 1293-1316, 2023.

[18] R. Phukan, N. Baruah, S. Sarma, and D. Konwar, "Parts-of-speech tagging for unknown words in assamese using viterbi algorithm," Indian Journal of Science and Technology, vol. 16, no. 1, pp. 53-59, 2023, doi: 10.17485/IJST/v16iSP2.8203.

[19] T. R. Jithendar, M. T. Devi, and G. Saritha, "DETERMINATION OF VITERBI PATH FOR 3 HIDDEN AND 5 OBSERVA-BLE STATES USING HIDDEN MARKOV MODEL," Reliability: Theory & Applications, vol. 19, no. 2 (78), pp. 509-515, 2024.

[20] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," Neural Computing and Applications, vol. 35, no. 23, pp. 17095-17112, 2023, doi: 10.1007/s00521-023-08568-z.

[21] S. Sun and H. Gao, "Meta-AdaM: An meta-learned adaptive optimizer with momentum for few-shot learning," Advances in Neural Information Processing Systems, vol. 36, no. 1, pp. 65441-65455, 2023.

[22] F. Mehmood, S. Ahmad, and T. K. Whangbo, "An efficient optimization technique for training deep neural networks," Mathematics, vol. 11, no. 6, pp. 1360-1381, 2023, doi: 10.3390/math11061360.