Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Analysis and evaluation of college entrance examination questions based on clustering algorithm

Yuqing Mo[1,✉]

[1] *Hunan College of Information, Changsha, Hunan, 410200, China*

ABSTRACT

This paper analyzes and evaluates high school examination questions based on machine learning. The study first introduces Bloom's classification method and constructs a categorized dataset of high school exam questions according to three steps of data collection, data annotation and data analysis. Then an automatic assessment model (WoBERT-CNN) based on WoBERT and Text-CNN is designed. The semantic similarity of word vector mapping is used to label the cases for determination, the improved WoBERT encoder is used to represent the text in word vectors, Text-CNN is used as a text classifier to extract the textual semantic features, and the features are integrated and screened, so as to realize the automatic classification of the cases in Bloom's taxonomy. Finally, based on the deep representation framework, the text information of the test questions is deeply mined and utilized to establish the relationship between the text of the test questions and the actual difficulty, and to realize the difficulty prediction of the test questions.The classification accuracy of the WoBERT-CNN model reaches more than 92%.The prediction error range of the H-MIDP model on the score rate of the test questions is between 1.3% and 3.2%, which is not too far from the real value. In conclusion, the automatic assessment model and difficulty prediction model designed in this paper can be applied in the analysis and evaluation of high school test questions, helping the high school test paper proposition and talent cultivation strategy.

Keywords: Machine learning; Bloom taxonomy, WoBERT-CNN, deep characterization, high school examination questions

1. Introduction

Teaching activities are not individual activities, and educators need to teach in an orderly manner under the unified guidance of the Ministry of Education with the curriculum standards as the core [1]. The current high school schedule has been reformed in terms of both the curriculum program and the subject curriculum standards, and a series of updates have been made to the subject curriculum

✉ Corresponding author.
E-mail address: lxfmyq@126.com (Y. Mo).
Received 20 February 2024; Revised 25 April 2024; Accepted 20 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-393

standards, including updating the knowledge points of teaching, selecting relevant subject content, advocating the focus on the major concepts of the subject, ensuring clarity of the curriculum structure and contextualization of the content and emphasizing a theme-oriented approach, so as to promote the development of the four core literacies [2-5]. In terms of academic quality, teaching is guided to pay more attention to the purpose of nurturing people and the development of students' core literacy [6]. Emphasis is placed on improving students' ability to comprehensively apply knowledge to solve practical problems, helping teachers and students to grasp the depth and breadth of teaching and learning, providing an important basis for the proposition of stage evaluations, academic level exams and promotion exams, and promoting the organic convergence of teaching, learning, and exams to form a synergy of nurturing [7-9]. With the introduction of the new curriculum for high school subject courses, the problems of how to interface teaching, learning and exams and how to grasp the focus of teaching have been troubling subject teachers.

The college entrance examination, as a national-type examination for selecting talents, carries the role of a bridge connecting basic education and higher education, and is also an important link in the landing of core literacy [10-11]. In this context, the selection and setting of the content of the college entrance examination is not only related to the quality of talent selection, but also has an important guiding role for basic education [12-13]. The core literacy embodied in the gaokao questions is becoming more and more obvious, and the study of the gaokao questions in the examination of the subject core literacy content, specific requirements and forms of expression and other examination characteristics, which can help front-line teachers to recognize the core literacy embodied in the gaokao questions, and effectively grasp the direction of the gaokao questions, which is of guiding significance for high school examination preparation and review [14-16]. In addition, making full use of modern teaching hardware equipment, optimizing the design of teaching strategies and perfecting the teaching process, strengthening the teaching of subject knowledge, giving students more targeted and efficient review strategies to learn the curriculum knowledge, and effectively implementing the cultivation of subject core literacy [17-19].

In this paper, we analyze the cognitive process dimension and knowledge dimension that can be achieved by the high school examination questions based on Bloom's classification method, and construct a high-quality classification dataset of high school examination questions based on the corresponding verb and noun seed libraries for the cognitive process dimension and knowledge dimension of the case dataset. The constructed high school exam question dataset is used for Bloom classification using machine learning approach by combining BERT pre-trained language model with Text-CNN text classification model, and WoBERT-CNN automatic assessment model is proposed. Then CNN-based difficulty prediction model (C-MIDP) and RNN-based difficulty prediction model (R-MIDP) are proposed, as well as H-MIDP difficulty prediction model is proposed by combining CNN and RNN, and the model application effect is verified through experimental analysis.

2. Method

Machine learning is the study of improving algorithms through experience or data, aiming at algorithms that allow machines to learn patterns from large amounts of historical data, automatically discover patterns and use them for prediction [20]. In other words, machine learning means that a machine learns from data, and the more data it processes, the more accurate its predictions will be. For the analysis and assessment of high school exam questions, the study proposes Bloom classification criteria for test questions, annotates the position of the two-dimensional matrix of Bloom classification method that can be achieved by the test questions, and constructs the classification dataset of high school exam questions. Machine learning technology is used to construct an automatic classification model for high school examination questions. And based on the deep representation framework, the textual information of the test questions is deeply mined and utilized, and then combined with a large number of students' answer data, the relationship between the text of the test

questions and the actual difficulty is established, so as to realize the difficulty assessment of the test questions.

2.1. Construction of the dataset of high school exam questions

In this paper, GAOKAO-Bench is chosen as the main source of the dataset. The dataset contains high school exam questions from 2010 to 2022, and the six dimensions in the cognitive process of Bloom's taxonomy as well as the four knowledge dimensions are used as the case labels to form the annotated canonical dataset of high school exam questions. The preparatory work includes collecting the teaching materials of the high school exam questions, including textbooks and lesson plans, prearranging them into the form of data cases, constructing the Bloom taxonomy control two-dimensional table corresponding to the test questions, and asking the relevant experts of the high school exams to carry out the content extraction and labeling according to the two-dimensional table, and ultimately forming the professional dataset of the high school exam questions annotated by hand. The construction of the dataset includes three steps: data collection, data labeling and analysis.

2.1.1. Data acquisition. The test question data collection was selected from the GAOKAO-Bench dataset, which covers all subject questions of the Chinese college entrance examination from 2010 to 2022, totaling 2,811 test questions. Among them, there are 1781 objective questions and 1030 subjective questions. In this paper, the lesson plans of the 2020 new college entrance examination math paper I are used as a case dataset to conduct Bloom classification research on them.

2.1.2. Data annotation. The manual annotation of the case dataset was done collaboratively by three master's degree students and two mathematical experts, in which the main annotation task was equally distributed among the three graduate students, each of whom submitted the completed annotation cases to the other for review, and if there was any disagreement, the cases were submitted to the third person for adjudication, and in the event that none of the three people were able to make a decision, the mathematical experts were asked to give assistance in adjudicating the cases as a way of ensuring that the annotation task was accurate and reliable. . To facilitate the adjudication of the cognitive process dimension of Bloom's taxonomy based on the verbs in the cases. In this paper, we have compiled the Bloom taxonomy verb dictionary about the case dataset, which contains eight subcategories under the six cognitive process dimensions of Bloom's taxonomy, which are identify, recall (memory dimension), explain, exemplify, transform (comprehension dimension), perform (application dimension), differentiate (analyze dimension), check (evaluate dimension), and (create dimension).The 2020 new Gao Gao Mathematics Ⅰ The knowledge dimension labels of the paper contain four types, namely “Geometric and Algebraic Knowledge”, “Functional Knowledge”, “Preparatory Knowledge”, and “Probability and Statistics Knowledge”.

2.1.3. Data analysis. In this paper, 2655 data were finally labeled to count the length of the text in the categorical dataset of the 2020 New College Entrance Examination Mathematics Paper I test questions. Taking 50 as the step size and setting different thresholds, we analyze the proportion of the number of samples whose sample text lengths are less than the threshold in the total dataset of the categorical dataset of the test questions of the new college entrance examination mathematics paper I in 2020, and the results are shown in Fig. 1. The data in the dataset with sample text length in the range of 0~100 accounts for about 44.32% of the total dataset, and with the increasing length threshold, the proportion of data smaller than the threshold increases more and more slowly. When the threshold is set to 500, the data with sample text length less than 500 accounts for 99.77% of the total dataset. Increasing the length threshold further, the proportion can no longer be significantly improved. Therefore, through the above analysis, the maximum input length of text is specified as 510 in this study, and the first 510 characters are saved for samples with a length greater than 510.

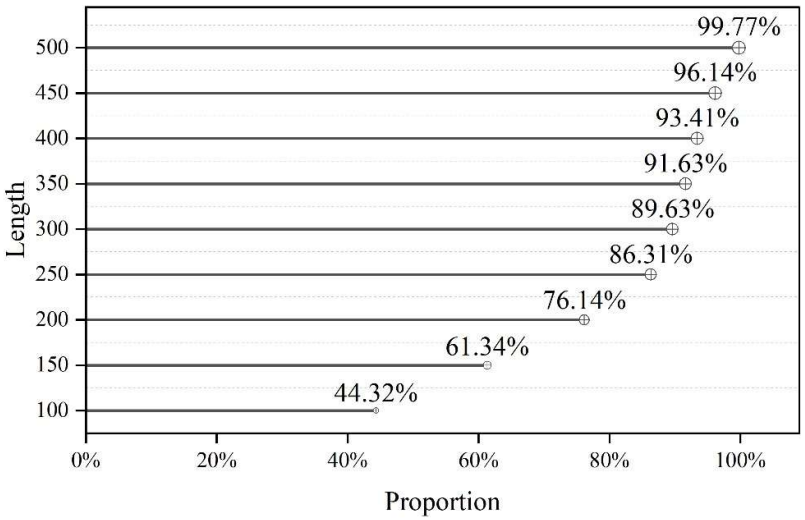

Fig. 1. The classification data set text length ratio

In the 2020 New College Entrance Examination Mathematics I. paper test questions in the Bloom classification two-dimensional matrix are listed in Table 1, in the knowledge dimension, the test questions in "geometry and algebra knowledge" are the most, accounting for 61.02% of the total, followed by the test questions in "function knowledge", accounting for 27.96%. Since the symbols, numbers, letters, events, places, people, times and other details in “Preparatory knowledge” were already embedded in other knowledge in the 2020 NSS Mathematics Paper I questions, the line “Preparatory knowledge” was deleted. In the cognitive process dimension, the most questions were at the comprehension level, accounting for 31.56%. Questions at the assessment level were the least, accounting for 4.03%.

Table 1. Distribution of cognitive process dimension and knowledge dimension

Knowledge dimension	Cognitive process dimension						Total	Percentage/%
	Memory	Understand	Applied	Analysis	Evaluate	Created		
Geometry and algebra	163	367	9	169	1	33	742	27.96%
Functional knowledge	65	312	412	405	106	320	1620	61.02%
Probability and statistics	61	159	21	12	0	40	293	11.02%
Total	289	838	442	586	107	393	2655	100%
Percentage/%	10.89%	31.56%	16.65%	22.07%	4.03%	0.148	100%	-

2.2. WoBERT-CNN Automatic Evaluation Modeling

The introduction of machine learning in the Bloom taxonomy [21] labeling classification task for the dataset can reduce the workload of manual labeling for the case dataset at a later stage and improve the quality and efficiency of the dataset construction. In this paper, the constructed dataset of high school exam questions is used for classification experiments in two dimensions of Bloom taxonomy using a machine learning approach that combines the BERT pre-trained language model [22] with the TextCNN text classification model [23], and proposes the WoBERT-CNN auto-assessment model. The model is used to analyze the high school exam questions from the cognitive process dimension, which in turn helps educators to reduce the difficulty of analyzing students' learning outcomes. The main framework of the model is shown in Figure 2. It mainly contains word embedding layer, text convolution layer, pooling layer and fully connected output layer.

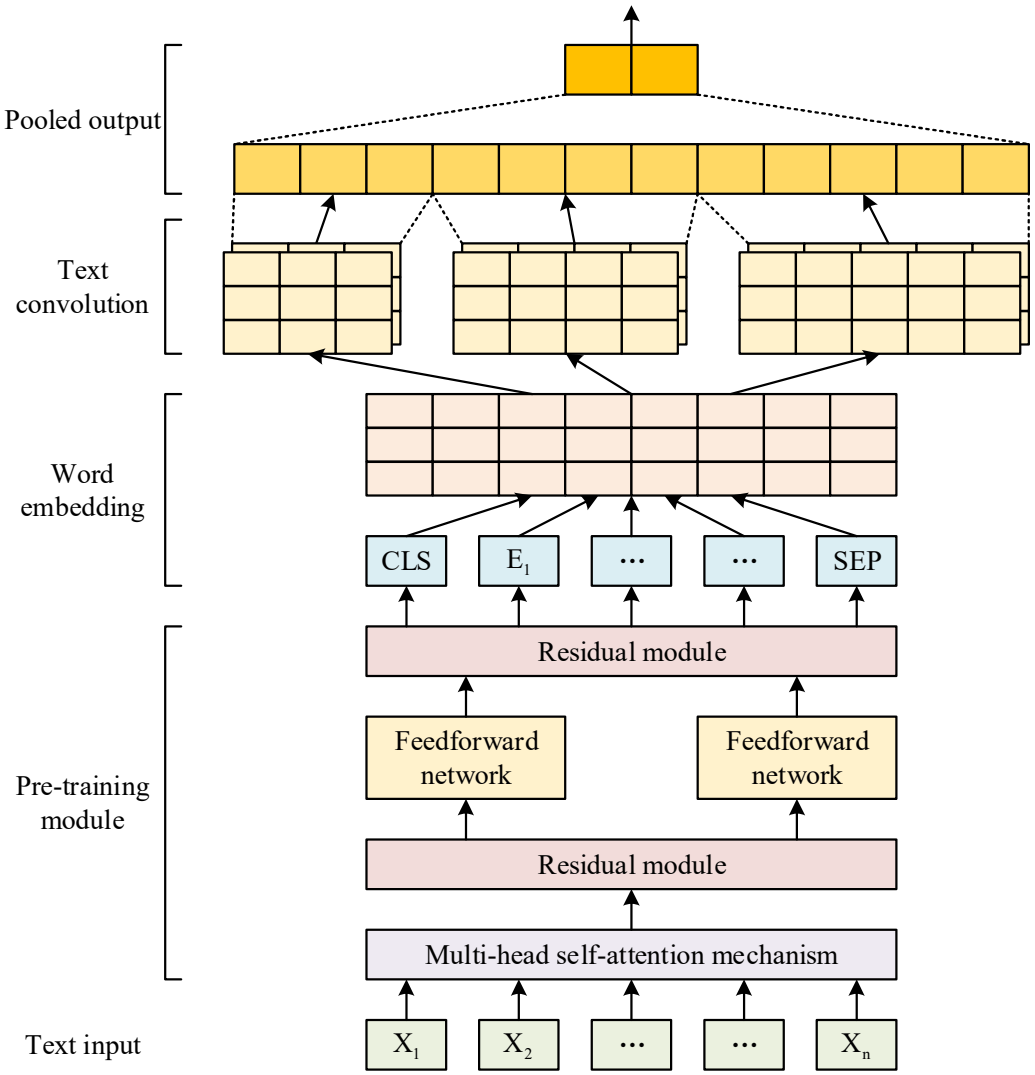

Fig. 2. WoBERT-CNN Classification model

2.2.1. Word Embedding Layer. The main role of the word embedding layer is to vectorize the input text using a vectorized representation based on the pre-trained language model WoBERT. BERT is a model structure with a multi-layer Transformer encoder, which can better capture the bi-directional contextual semantics of the text due to its bi-directional coding structure and the use of a multi-head attention mechanism to fuse the information.

The multi-head self-attention mechanism is to merge multiple sets of self-attention units so that each self-attention unit can be projected into the same dimensional space so that the encoder can better capture the interaction information in each dimension. The attention mechanism is the main constituent element of multi-head self-attention, which is described as shown in Equation (1):

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where Q,K,V represent query, key and value vector matrices, respectively.

Multihead self-attention combines multiple self-attention units. First, the input X is passed to h individual identical self-attention units, and h output matrices Z are computed, h being the number of Heads in Multi-Head Attention. Next, the h output matrices $Z_1 - Z_h$ are spliced and input to a Linear layer to obtain the final output Z. The computational formula is described as shown in equation (2).

$$Multi-Head(Q,K,V)=Concat(H_1,H_2,...,H_h) \tag{2}$$

Where Multi-Head represents the multi-head attention mechanism, Concat function represents the splicing operation and H is the number of Heads. The feed-forward propagation network layer is located after the self-attention unit in the encoder, and its purpose is to enhance the fitting ability of the self-attention unit and to make the parameter dimensions constant through the feed-forward network layer. This layer consists of two linear transformations, and there is also a ReLU activation function between the two transformations. The formula is shown in equation (3).

$$FFN(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{3}$$

where $\max$ denotes the ReLU activation function and x denotes the input parameters.

2.2.2. Text Convolution Layer. In this paper, Text-CNN is used as a text classification model, the model network structure is concise, the number of parameters is small, and it can make full use of word vectors containing semantic information for feature extraction and classification, for the constructed case dataset, the model is able to utilize the semantic features in the case text for the two dimensions of Bloom's taxonomy. In this paper, for the case dataset of 2020 New College Entrance Examination Mathematics Paper I, after obtaining the trained word vectors by pre-training the language model, the TextCNN text classification model is utilized to conduct multi-classification experiments, and the structure of this text input TextCNN model is shown in Fig. 3.

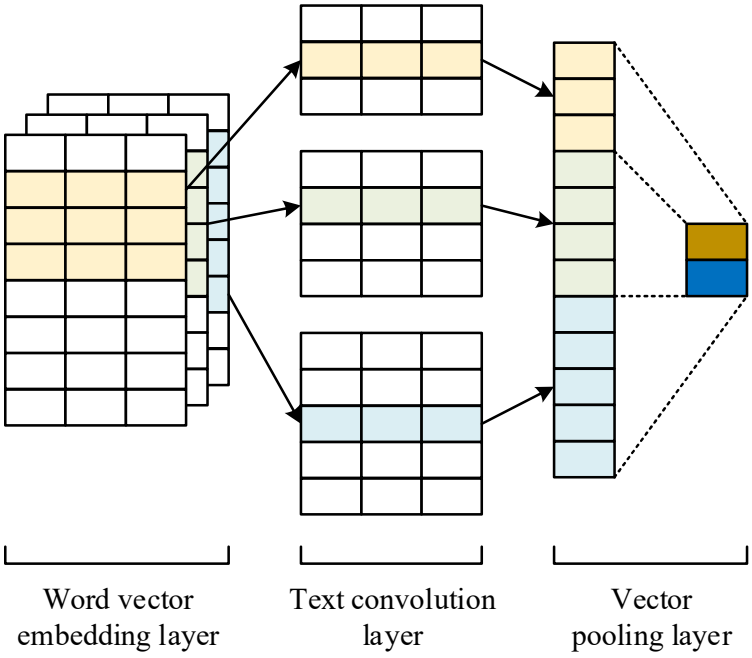

Fig. 3. Text-CNN Language model

After obtaining the pre-trained word vector matrix, the text convolutional layer will realize the local feature extraction of the input text by connecting the input and output neurons, and connect the local features in the text for the classification task of the case text. The sharing of parameters between the same layers of the text convolutional layer effectively reduces the complexity of the convolutional neural network model. The text convolution operation is shown in Figure 4. Where the word vector matrix on the left represents the feature matrix X of a certain data, and the size of the matrix is 5*5. The feature extraction operation of the text is realized by a 3*3 convolution kernel W to obtain a feature matrix of size 3*3.

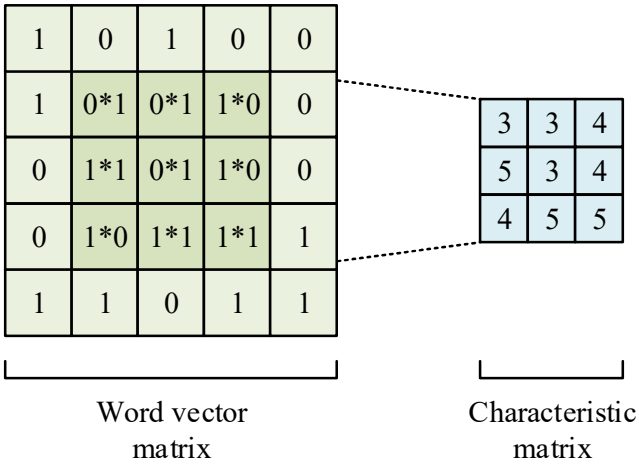

Fig. 4. Text convolution

The width of the convolution kernel in the text convolution layer is the same as the dimension of the word vector matrix I, i.e., the word vector dimension k, and the convolution kernel W with a sliding window in the vertical direction $A[i:i+h-1]$ does a convolution operation on the input $n \times k$ -dimensional word vector embedding matrix to get the feature c_i expression as shown in Eq. (4):

$$c_i = f(W \cdot A_{i:i+h-1} + b) \tag{4}$$

where h represents the number of words in the sliding window, $f(\cdot)$ is the nonlinear activation function tanh, W denotes the $h \times k$ -dimensional weight matrix corresponding to the convolution kernel, $Ai:i+h-1$ denotes the sliding window of size $h \times k$ dimensions assembled from rows i to $i+h-1$ of the input weight matrix, and b is the bias parameter on the superposition.

2.2.3. Pooling layer. After the text convolution operation on the feature vectors, in order to ensure that the size of the feature maps obtained from different sizes is consistent, the maximum pooling function is used to extract the maximum value of the feature vectors in each sliding window to represent the feature, and the extracted features are cascaded to obtain the final vector representation of the pooling layer.

The maximum pooling expression is shown below:

$$\hat{c} = \max(c) \tag{5}$$

where c is the feature vector in the current sliding window and $\hat{c}$ denotes the maximum value in the feature vector.

2.2.4. Fully Connected Output Layer. The fully-connected output layer mainly plays the role of a classifier, and after the pooling layer operation to obtain the word vector representation of the case text, the fully-connected layer is connected to output the probability of each case text belonging to the cognitive process dimension and the knowledge dimension of Bloom's taxonomy, and the probability distribution is obtained by using the nonlinear activation function Softmax to normalize the probability. The reason for using the Softmax function is that this activation function no longer uniquely determines a certain maximum value, but instead assigns to each output a probability value indicating the likelihood that the corresponding input text belongs to a category.The Softmax function is defined as shown in Equation (6):

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}} \tag{6}$$

Where z_i is the output value of the i th node, and C is the number of output nodes, i.e., the number of categories of classification. Using Softmax function can convert the output probability value of multi-categorization into a probability distribution ranging between [0, 1], which is finally converted into classification labels of two dimensions of Bloom's classification according to the probability distribution law.

2.3. Deep neural network-based test question difficulty assessment

This chapter proposes the characterization models C-MIDP, R-MIDP and H-MIDP for test texts to assess the difficulty of high school test questions.

2.3.1. Problem definition. The formal definitions for the concepts of exam, test, and score rate are as follows:

Define $Q = Q_1, Q_2, \cdots, Q_n$ as the set of test questions and $T = T_1, T_2, \cdots, T_m$ as the set of high school papers. $T_i = \{\tilde{Q}_i, \tilde{R}_i\}$, where $\tilde{Q}_i$ and $\tilde{R}_i$ are the set of exercises in the exam T_i and the set of scores, respectively. $\tilde{Q}_i = \{\tilde{Q}_{i1}, \tilde{Q}_{i2}, \cdots, \tilde{Q}_{im_i}\}$, $\tilde{Q}_{ij} \in Q$, is the number of questions in the exam T_i of the number of questions in the test. $\tilde{R}_i = \{\tilde{R}_{i1}, \tilde{R}_{i2}, \cdots, \tilde{R}_{im_i}\}$, $\tilde{R}_{ij}$ is the examination T_i The score rate of the j th test question in the exam is used as the true value of the difficulty of the test question in the exam T_i. The score rate is calculated as follows:

$$R_{ij} = \frac{\text{Sum of scores on question } j \text{ in exam } i}{\text{Answer record of question j in exam i} \times \text{Answer record of question } j \text{ in exam } i} \tag{7}$$

Given a collection of high school exam questions Q and a collection of high school exam records T, where Q contains the text of each test question and T contains the questions and corresponding scores for each exam, the goal is to model high school exams such that the predicted value of the difficulty of the questions can be obtained by inputting the features of the questions into the model.

The notation involved in the problem and the corresponding descriptions are shown below:

Q - the full set of test questions.

Q_i - test questions i.

R_i - the score rate of test question i.

P_i - predicted difficulty (score rate) of test question i.

T - the set of high school exams.

T_i - exams in a subject i.

$\tilde{Q}_i$ -the set of test questions in a subject exam i.

$\tilde{Q}_{ij}$ - the set of test questions j in a subject exam i.

$\tilde{R}_i$ - the set of scores in a subject test i.

$\tilde{R}_{ij}$ - the score rate of test question j in a subject exam i.

X_q - the characteristics of test question q.

2.3.2. Model structure. The three models proposed in this paper accept test question features as input and the output is the predicted difficulty of the test questions. The test question features are obtained by splicing the word vectors of text characters in the following steps.

1) Obtain the characterization vectors of Chinese words or English characters of the test questions. Using the word2vec method, take the 2020 New College Entrance Examination Mathematics Paper I test question text collection as the training data, and train to get the word vector $\vec{w}_i \in R^{d_0 \times 1}$ for each word or character in the text database. Taking the trained word vectors as the representation vectors

of the corresponding words or characters has a lower dimension than the one-hot representation and can maintain certain semantic features.

2) Construct the initial representation vector of the test question. Segment the test question and remove the deactivated words, and replace the remaining words or characters with the corresponding word vectors in the order of the original text to obtain the initial representation vector of the test question $X_q^{'} \in R^{d_0 \times n}$, where n is the number of word vectors.

3) Modify the length of $X_q^{'}$ to get the final representation vector of test questions X_q. Since the input features of the model need to be of fixed length, choose the appropriate length N (N is the number of word vectors, set N =600 in the experiment), and fill it with $N-n$ zero vectors if $n<N$, and on the contrary, delete the last $n-N$ word vectors if $n>N$. The final X_q is used as the characterization vector of the test question.

After the text of the test question is converted into vector features, it is input into the model for semantic understanding, and the model structure is described specifically in the following.

1) C-MIDP model.The C-MIDP model is based on CNN, and the multilayer convolutional and pooling layers used can learn test question information from different levels. The model includes an input layer, two convolutional and Max Pooling layers, a fully connected layer and an output layer. The input layer accepts the test question features $X_q \in R^{d_0 \times N}$. X_q performs a convolution operation in the first convolutional layer through dl $k \times 1$ convolutional kernels, outputting a hidden layer of dl channels.

Specifically, given the input $X_q = \{\vec{w}_1, \vec{w}_2, \cdots, \vec{w}_N\}$, $\vec{w}_1 \in R^{d_0 \times 1}$, the convolution to obtain the The hidden layer $H^c = \{\vec{h}_1^c, \vec{h}_2^c, \cdots, \vec{h}_{N+k-1}^c\} \in R^{d_1 \times (N+k-1)}$, where:

$$\vec{h}_j^c = ReLU\left(\sum_{j=1}^{d_0} G \cdot [w_{i-k+1}^j \cdots w_i^j] + b\right) \tag{8}$$

where $G \in R^{d_1 \times k}$, i.e., d_1 a convolution kernel of length k. w_i^j denotes the value of the j -dimensional element of $\vec{w}_i$: $b \in R_1^d \times 1$ is the bias term: $ReLU(x) = \max(0, x)$ is the nonlinear activation function.

The output of the first convolutional layer h^c is passed through a p-max pooling layer to pick out the most important information in it to get a new hidden layer $H^{cp} = \{\vec{h}_1^{cp}, \cdots, \vec{h}_{(N+k-1)/p}^{cp}\}$, where:

$$\vec{h}_i^{cp} = \left[\max\begin{bmatrix} h_{i\times p-p+1,1}^c \\ \cdots \\ h_{itimesp,1}^c \end{bmatrix}, \cdots, \max\begin{bmatrix} h_{i\times p-p+1,d_1}^c \\ \cdots \\ h_{itimesp,d_1}^c \end{bmatrix}\right] \tag{9}$$

The h^{cp} again goes through a convolutional layer and a p-max pooling layer, which operates similarly to the first layer. Let the number of convolution kernels in the second convolutional layer be d_2, and take the window size of the second Max Pooling layer to be $(N+k-1)/p$, at which point the output hidden layer $H_2^c \in R^{d_2 \times 1}$ is transposed to pass through a fully-connected layer, outputting a prediction of the difficulty of the test question P_q.

2) R-MIDP model.In addition to obtaining localized information, the sequential semantic and logical information of the text is also very important for understanding the test question. For example, a number in a formula may not contain much information by itself, but may exhibit important semantics if it is associated with a number of characters preceding it. Based on this, this section proposes the R-MIDP model, which is based on RNN and utilizes the Cell module in RNN to save historical information and learn the sequential semantic or logical information of the text of the test questions. Specifically, the R-MIDP model is a two-way LSTM network structure, which can learn the semantic logic of test questions from both forward and reverse directions in the process of understanding test questions to make the semantics more complete.

The R-MIDP model consists of an input layer, a bidirectional LSTM hidden layer, a Max Pooling layer, a fully connected layer, and an output layer. The input layer accepts the trial features $X_q \in \{\vec{x}_1, \cdots, \vec{x}_N\} \in R^{N \times d_0}$, and $\vec{x}_i = \vec{w}_i^T$. The number of LSTM layer cells is d. After single layer LSTM to get the hidden layer $H^r = \{\vec{h}_1^r, \cdots, \vec{h}_d^r\} \in R^{N \times d}$, $\vec{h}_j^r = (h_{j,1}^r, \cdots, h_{j,N}^r)$, where $h_{j,t}^r$ is obtained from the following LSTM formula:

$$i_t = \sigma(W_{j,ii} x_t + b_{j,ii} + W_{j,hi} h_{j,(t-1)}^r + b_{j,hi}) \tag{10}$$

$$f_t = \sigma(W_{j,if} x_t + b_{j,if} + W_{j,hf} h_{j,(t-1)}^r + b_{j,hf}) \tag{11}$$

$$g_t = tanh(W_{j,ig} x_t + b_{j,ig} + W_{j,hc} h_{j,(t-1)}^r + b_{j,hg}) \tag{12}$$

$$o_t = \sigma(W_{j,io} x_t + b_{j,io} + W_{j,ho} h_{j,(t-1)}^r + b_{j,ho}) \tag{13}$$

$$c_t = f_t * c_{(t-1)} + i_t * g_t \tag{14}$$

$$h_{j,t}^r = o_t * \tanh(c_t) \tag{15}$$

H' goes through a p-max pooling layer to obtain a new hidden layer $H^{rp} = \{h_1^{rp}, \cdots, h_d^{rp}\}$, where:

$$h_j^{rp} = \max(h_{j,1}^r, \cdots, h_{j,N}^r) \tag{16}$$

H^{rp} goes through a fully connected layer and outputs the predicted value of the difficulty of the test questions P_q.

3) H-MIDP model.Further, this paper combines the advantages of two models, C-MIDP and R-MIDP, and proposes a hybrid model, H-MIDP, with a view to modeling the locally important semantic and sequential logical information of the test question text effectively at the same time. The first half of the model is divided into two parallel parts, one of which is the same as C-MIDP, in which the input features are subjected to two convolutional layers and Max Pooling layer to obtain the hidden layer value H_2^c. The other part is the same as R-MIDP, the input features are passed through LSTM and Max Pooling layer to get the hidden layer value H^{rp}, and H_2^c and H^{rp} are spliced together to get $H \in R^{1 \times (d_2 + d)}$, and the test question is obtained through the two layers of fully connected layers. The predicted value of difficulty P_q.

2.3.3. Model training. This section describes the learning approach of the proposed model. First, a supervised training approach is followed to learn the model in large-scale historical exam data. Specifically, in the training data, i.e., the historical exam data, the difficulty prediction values are obtained by modeling the text of the test questions using the three models mentioned above, using the test score rate as a label and using the formula shown in Eq. (17) as the objective function of the model:

$$L(T) = \sum_q (P_q - R_q)^2 \tag{17}$$

where T is the entire math test training set, and P_q and R_q are the predicted difficulty and actual score rate of test question q, respectively. This approach is often performed on a test-question-by-test-question basis when calculating test-question scores, and its training is in fact not differentiated among different student groups. However, in practice, different student groups usually make the computational difficulty of test questions behave differently, specifically, the score rate is not comparable due to the difference in student groups in different exams, different schools, and other scopes.

In order to be able to eliminate this effect, the three models in this paper adopt context-related training, using equation (18) as the model's loss function:

$$L(T)=\sum_{(T_t,Q_i,Q_j)}((P_{ti}-P_{tj})-(\tilde{R}_{ti}-\tilde{R}_{tj}))^2 \tag{18}$$

Where T_t denotes the context range t, and P_n and P_{ij} refer to the actual difficulty (score rate) of the test questions Q_i and Q_j in the $contextT_i$, respectively.

Using such a loss function removes the variability of different groups of students in different contexts and captures the commonalities, allowing the trained model to predict the true difficulty of the test questions (the difficulty for the entire population of students involved in all the answer logs, rather than for a particular group of students in one of the exams). Such difficulty values are then more accurate and interpretable.

3. Results and Discussion

The study takes the 2020 New College Entrance Examination Mathematics I paper as an example, classifies it based on the WoBERT-CNN automatic assessment model, and evaluates the difficulty of the test questions based on deep neural networks.

3.1. Experimental analysis of the classification of high school exam questions

3.1.1. Metrics. Test question cognitive process dimension and knowledge dimension classification belongs to multi-categorization problems, in order to calculate the precision rate, recall rate and Fl value of the classification model on different categories, this paper selects the weighted precision rate, weighted recall rate and weighted F1 value as the model prediction evaluation indexes:

$$W-P=\sum_{i=1}^{K}\left(P_i\times\frac{n_i}{K}\right) \tag{19}$$

$$W-R=\sum_{i=1}^{K}\left(R_i\times\frac{n_i}{K}\right) \tag{20}$$

$$W-F1=\sum_{i=1}^{K}\left(\frac{2\times P_i\times R_i}{P_i+R_i}\times\frac{n_i}{K}\right) \tag{21}$$

where K is the number of categories, P_i is the precision rate, R_i is the recall rate, and n_i is the number of samples in the ith category.

3.1.2. Data and design. In order to verify the effectiveness of the proposed model, this paper introduces three classical classification models for comparison experiments.The three classical classification models include CNN, Text-CNN and BERT model. The training set and test set of this experimental dataset are divided in the ratio of 7:3, and this paper adopts the character-based method to preprocess the data. By analyzing the data in subsection 2.1.3, the test text is short-filled and long-cut in the experiment, and the length of each sentence is treated as 510.

3.1.3. Analysis of experimental results. The experimental results of WoBERT-CNN and the baseline method under the cognitive process dimension are shown in Fig. 5.The weighted precision, weighted recall, and weighted Fl values of the WoBERT-CNN model are 0.852, 0.839, and 0.868, respectively, which are higher than those of the baseline model.

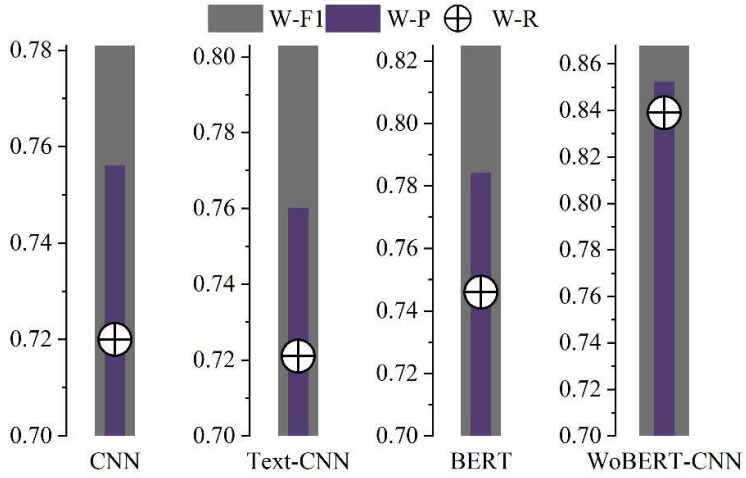

Fig. 5. Results of cognitive process dimension classification of test questions

The experimental results of WoBERT-CNN and baseline methods under knowledge dimension are shown in Fig. 6.The weighted precision, weighted recall and weighted Fl values of WoBERT-CNN model are more than 0.9, while the values of the three metrics of the other baseline models are below 0.9.

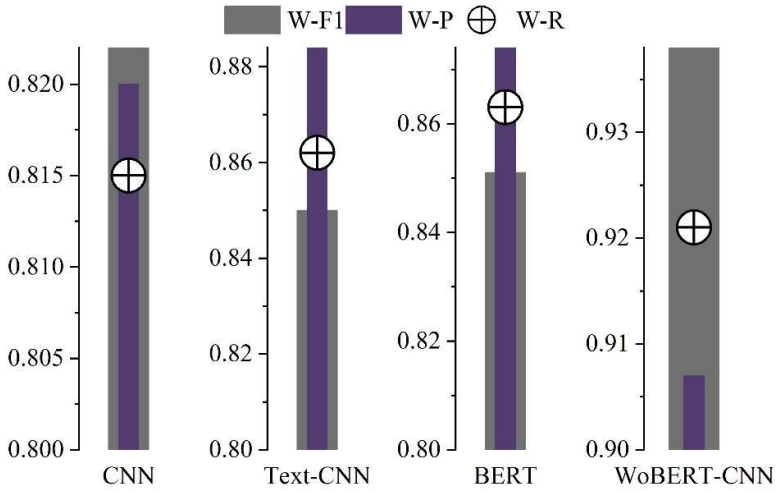

Fig. 6. Results of knowledge dimension classification of test questions

The error statistics of the classification results at all levels of the cognitive process dimension of the test set are shown in Table 2. The machine classification ignores the contextual context of the test questions to a certain extent, and at the same time, the course experts annotate the test questions with a certain degree of subjective consciousness, but these errors are within the acceptable range, and the model classification accuracy rate is above 92%. It shows that the results of analyzing the high school test questions based on the WoBERT-CNN model have practical reference value.

Table 2. The cognitive process dimension is not classified

Grade	Test number	Accuracy/%	Error	Memory	Understand	Applied	Analysis	Evaluate	Created
Memory	41	92.68	3	-	1	0	0	1	1
Understand	263	97.72	6	3	-	0	1	0	2
Applied	152	97.37	4	0	1	-	2	1	0
Analysis	169	94.08	10	5	1	0	-	0	4
Evaluate	33	96.97	1	0	0	1	0	-	0
Created	116	94.83	6	2	0	1	3	0	-

3.2. Analysis of the difficulty assessment of high school exam questions

3.2.1. Assessment of indicators. In order to measure whether the difficulty assessment model proposed in this paper can effectively predict the difficulty of test questions, three indicators, Pearson correlation coefficient (PCC), consistency (DOA) and root mean square error (RMSE), are used in the experiments to evaluate the prediction effect of the model, and the computational form of each indicator is shown below:

$$PCC = \frac{\sum_{i=1}^{u_v}(P_{iv} - \overline{P}_v)(Z_{iv} - \overline{Z}_v)}{\sqrt{\sum_{i=1}^{u_v}(P_{iv} - \overline{P}_v)^2}\sqrt{\sum_{i=1}^{u_v}(Z_{iv} - \overline{Z}_v)^2}} \tag{22}$$

Where u_v is the number of questions in the v nd exam, $P_{iv}, \overline{P}_v$ is the predicted and average predicted difficulty of question i in that exam, and $Z_{iv}, \overline{Z}_v$ is the actual and average actual difficulty of question i in that exam.

$$DOA = \sum_{(C,L)\in u_v}\left(\frac{\sigma[(P_{iC}, P_{iL}) \wedge (Z_{iC}, Z_{iL})]}{\sigma(Z_{iC}, Z_{iL})}\right) \tag{23}$$

$$\sigma(x, y) = \begin{cases} 1, x > y \\ 0, x \le y \end{cases} \tag{24}$$

P_{iC} and P_{iL} are the predicted difficulty of the topic C, L at the time of the test, and Z_{iC} and Z_{iL} are the actual difficulty of the topic C, L, and $\sigma(x, y)$ is defined as shown in Eq. (24). The evaluation index measures the correlation between the actual difficulty of the questions at the time of testing and the difficulty predicted by the model.PCC takes the value in the range of [-1, 1], and the larger the absolute value indicates the higher the correlation. DOA is to evaluate the accuracy of the relative comparison of the difficulty prediction values of different test questions during the test, the value range is between 0-1, the larger the DOA value indicates that the difficulty prediction value of the test questions is more accurate. RMSE is used to measure the gap between the predicted difficulty and the actual difficulty, the value close to zero indicates a better performance.

In summary, the smaller the value of RMSE, the better, while the larger the values of PCC and DOA within their range of values, the better the model performance.

3.2.2. Model comparison. In this section, C-MIDP, R-MIDP and H-MIDP models are compared and analyzed to observe the performance of the pairs of models in predicting the difficulty of high school exam questions. The results of the model comparison are shown in Fig. 7.The H-MIDP model consistently outperforms the C-MIDP and R-MIDP models in the three evaluation metrics of RMSE, PCC

and DOC. This indicates that the hybrid model H-MIDP is more superior in predicting the difficulty of high school exam questions.

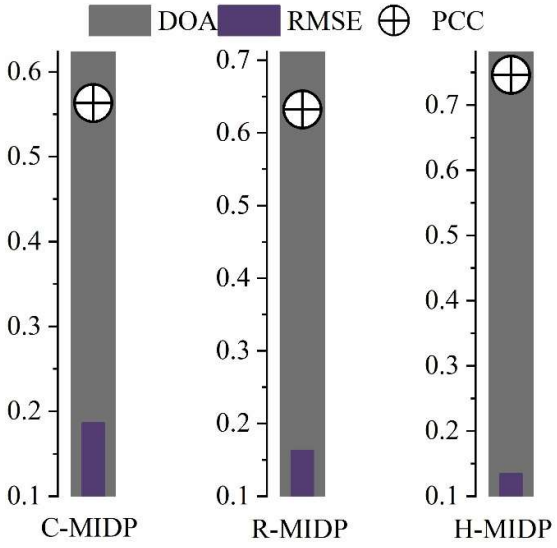

Fig. 7. Model comparison results

3.2.3. Assessment analysis. The H-MIDP model was used to predict the difficulty of "Knowledge of Geometry and Algebra", (D1) "Knowledge of Functions" (D2), and "Knowledge of Probability and Statistics" of the 2020 NSS Mathematics I paper (D3) were predicted in terms of difficulty. The comparison between the predicted and actual scores of the model is shown in Figure 8.The prediction errors of the H-MIDP model for the three knowledge dimensions of the 2020 Mathematics I paper of the new college entrance examination are 1.4%, 3.2% and 1.3%, respectively, which are closer to the real values. It is closer to the real value, indicating that the model can be used for the prediction of the high school exam questions.

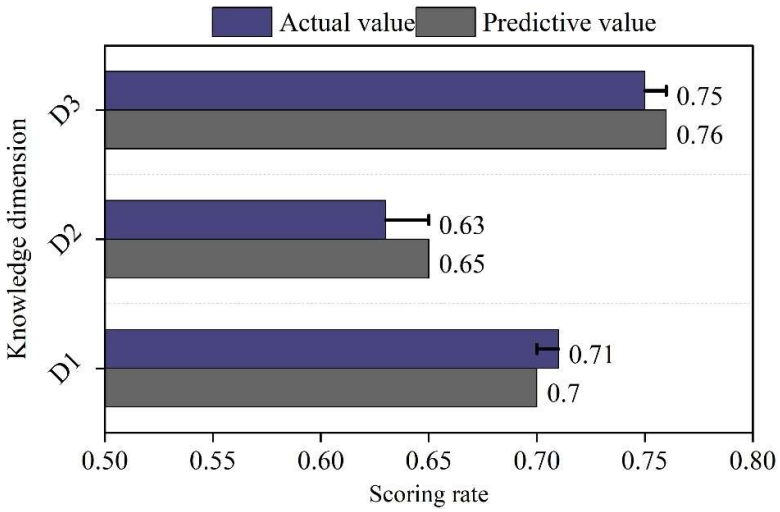

Fig. 8. The ratio of the model prediction to the actual score ratio

4. Conclusion

This paper proposes an automatic classification method of test questions based on Bloom classification method, and realizes automatic prediction of test question difficulty based on deep neural network model. Taking the 2020 new college entrance examination math paper Ⅰ questions

as an example, Bloom classification method divides the paper into “geometry and algebra knowledge”, “function knowledge” and “probability and statistics knowledge”. 3 knowledge dimensions. A WoBERT-CNN model combining WoBERT and Text-CNN is used to classify the questions, and the classification accuracy is over 92%. The scores of these 3 knowledge dimensions were predicted based on the H-MIDP model, and the prediction errors were 1.4%, 3.2%, and 1.3%, respectively.The classification errors of the WoBERT-CNN model and the prediction errors of the H-MIDP model were within the acceptable range. The method of this paper provides strong support for the analysis and evaluation of high school exam questions.

References

[1] Fazlina, A. (2018). An analysis of college entrance test. English Education Journal, 9(2), 192-215.

[2] Zheng, M., & Ruan, C. (2024). The new college entrance examination reform investigation and research on comprehensive quality evaluation. Heliyon, 10(19).

[3] Huang, J. (2020). Operation mechanism and evaluation of “county high school education model” in the context of Chinese college entrance examination system. Science Insights Education Frontiers, 7(2), 879-891.

[4] Mengna, Z., & Chengwu, R. (2024). Investigation and research on comprehensive quality assessment in the reform of new college entrance examination. South African Journal of Education, 44(1), S1-S18.

[5] Lin, Y., & Weatherly, K. I. C. H. (2024). Standardizing “Yikao”: An analysis of the reform in Arts College Entrance Examination (ACEE) in China in 2024. International Journal of Music Education, 02557614241269062.

[6] Yao, J. X., & Guo, Y. Y. (2018). Core competences and scientific literacy: The recent reform of the school science curriculum in China. International Journal of Science Education, 40(15), 1913-1933.

[7] Yao, Y., Zhang, Z., Cui, H., Ren, T., & Xiao, J. (2019). The influence of student abilities and high school on student growth: A case study of chinese national college entrance exam. IEEE Access, 7, 148254-148264.

[8] Huang, S. (2025). Analysis of the Influence of College Entrance Examination System Reform on Educational Equity. Journal of Social Science Humanities and Literature, 8(1), 1-7.

[9] Wang, Z. (2024, November). Analysis of the “trinity” Enrollment Status and Optimization Measures under the Background of the New College Entrance Examination. In 2024 4th International Conference on Public Relations and Social Sciences (ICPRSS 2024) (pp. 281-290). Atlantis Press.

[10] Hong, D. S., Bae, Y., & Wu, Y. F. (2019). Alignment between National College Entrance Examinations and Mathematics Curriculum Standards: A Comparative Analysis. Research in Mathematical Education, 22(3), 153-174.

[11] Allen, D. (2020). Proposing change in university entrance examinations: A tale of two metaphors. TEVAL-Shiken: A Journal of Language Testing and Evaluation in Japan, 24(2), 23-38.

[12] Talman, K., Vierula, J., Karihtala, T., Laakkonen, E., Engblom, J., & Haavisto, E. (2025). Development and Validity Evaluation of a National Digital Entrance Examination Test for Higher Education Student Selection. Higher Education Quarterly, 79(1), e70006.

[13] Zarei, R., Ojinejad, A. R., & Saremi, M. (2016). Analyzing University Entrance Examination Questions in the Field of Study of Mathematics-Physics. Journal of New Approaches in Educational Administration, 7(27), 139-162.

[14] Chandralekha, E., Jemin, V. M., Rama, P., & Prabakaran, K. (2023). Sentiment analysis of National Eligibility-cum Entrance Test on Twitter data using machine learning techniques. Advances in Science and

Technology, 124, 344-354.

[15] Li, Y. X., Miao, Y. N., & Lue, S. Y. (2017, April). A comprehensive analysis method of application in college entrance examination. In 2017 International Conference on Network and Information Systems for Computers (ICNISC) (pp. 131-134). IEEE.

[16] Keramati Nojedehsadat, M., & Gholami, A. (2023). Analysis of University Entrance Exam Questions and Candidates' Performance in Answering Them Based on Bloom's cognitive Levels. Pouyesh in Humanities Education, 9(30), 63-86.

[17] Smadi, A., Al-Qerem, A., Nabot, A., Jebreen, I., Aldweesh, A., Alauthman, M., ... & Alzghoul, M. B. (2023). Unlocking the potential of competency exam data with machine learning: improving higher education evaluation. Sustainability, 15(6), 5267.

[18] Testa, I., Capasso, G., Colantonio, A., Galano, S., Marzoli, I., Scotti di Uccio, U., & Serroni, G. (2020). Validation of university entrance tests through Rasch analysis. Rasch Measurement: Applications in Quantitative Educational Research, 99-124.

[19] Han, C., & Xiang, J. (2025). Alignment Analysis Between China College Entrance Examination Physics Test and Curriculum Standard Based on E-SEC Model. International Journal of Science and Mathematics Education, 23(1), 215-234.

[20] Nakai Tomoya,Tirou Coumarane & Prado Jérôme. (2024). From brain to education through machine learning: Predicting literacy and numeracy skills from neuroimaging data. Imaging Neuroscience,2,1-24.

[21] Rajesh Dontham,Savita. P. Patil,Manoj. A. Patil & Anupama. K. Ingale. (2016). Educating IT-Engineering Theory Courses Using BLOOMS Taxonomy(TC-BT) as a Tool and Practical Courses Using Magic of Making Mistakes (PC-MoMM):An Innovative Pedagogy. Journal of Engineering Education Transformations,29(3),75-84.

[22] Konstantinos I. Roumeliotis,Nikolaos D. Tselikas & Dimitrios K. Nasiopoulos. (2024). Leveraging Large Language Models in Tourism: A Comparative Study of the Latest GPT Omni Models and BERT NLP for Customer Review Classification and Sentiment Analysis. Information,15(12),792-792.

[23] Han Chengcheng,Lin Qiang,Man Zhengxing,Cao Yongchun & Wang Haijun. (2021). Automated Classification of Textual SPECT Diagnostic Reports with TextCNN Model. Journal of Physics: Conference Series,1792(1),012023-.