IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 22 July 2022, accepted 7 August 2022, date of publication 16 August 2022, date of current version 22 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3198682

RESEARCH ARTICLE

Can We Predict Student Performance Based on Tabular and Textual Data?

YUBIN QU[1,2], FANG LI[3], LONG LI[4], (Member, IEEE), XIANZHEN DOU[2], AND HONGMEI WANG[5]

[1]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[2]School of Information Engineering, Jiangsu College of Engineering and Technology, Nantong 226001, China
[3]School of Marxism, Jiangsu College of Engineering and Technology, Nantong 226001, China
[4]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[5]School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China

Corresponding author: Hongmei Wang (wanghongmei@just.edu.cn)

This work was supported in part by the Jiangsu Province Education Science "14th Five-Year Plan" under Project D/2021/01/133, in part by the Philosophy and Social Science Research Projects in Jiangsu under Grant 2020SJB0836, in part by the Nantong Science and Technology Project under Grant JC2021124, in part by the Guangxi Key Laboratory of Trusted Software under Grant kx202046, in part by the Scientific Research Projects of Jiangsu College of Engineering and Technology under Grant GYKY/2020/4, in part by the Research Project of Modern Educational Technology in Jiangsu Province under Grant 2021-R-94735, in part by the Special Project of China Higher Education Association under Grant 21SZYB23, in part by the Fifth Jiangsu Province Vocational Education Teaching Reform Research Project under Grant ZYB686, in part by the Special Foundation for Excellent Young Teachers and Principals Program of Jiangsu Province, and in part by the Qing Lan Project of Jiangsu Province.

ABSTRACT With the emergence of more new teaching systems, such as Massive Open Online Courses (MOOCs), massive amounts of data are constantly being collected. There is a huge value in these massive teaching data. However, the data, including both student behavior data and student comment data about the course, is not processed to discover models and paradigms which can be useful for school management. There is no multimodal dataset with tabular and textual data for educational data mining yet. We first collect a dataset that included student behavior data and course comments textual data. Then we fuse the student behavior data with course comments textual data to predict student performance, using a Transformer-based framework with a uniform vector representation. The empirical results of the collected dataset show the effectiveness of our proposed method. In terms of F1 and AUC the performance of our method improves by up to 3.33% and 4.37% respectively. We find that the uniform feature vector representation learned by our proposed method can indeed improve the classifier's performance, compared with existing works. Further, we validate our approach on an open dataset. The results of the empirical study show that our proposed method has a strong generalization capability. Moreover, we perform interpretability analysis using the SHapley Additive exPlanation (SHAP) method and find that text features have a more important influence on the classification model. This further illustrates that fusing text features can improve the performance of classification models.

INDEX TERMS Educational data mining, deep learning, multimodal, data fusion, random forest.

I. INTRODUCTION

Traditional educational institutions have accumulated much information about the student, including the student's school number, age, gender, etc. This data is usually stored in a relational database. This type of data is called tabular data. With the development of the mobile Internet in recent years, web-based online educational systems have flourished exponentially, thus providing multiple data sources with different granularity levels [1], [2]. With the emergence of more new teaching systems, such as MOOCs, massive amounts of data are constantly being collected. There is a huge value in these massive teaching data. However, this data is not being processed in time to discover models and paradigms useful for school management. In fact, the tension between the sheer size of data and knowledge discovery is a huge challenge

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojun Steven Li.

for educational institutions today [3]. The applications of data mining techniques for the specific data from educational environments are called educational data mining (EDM) [4]. EDM data comes from a wide variety of educational systems, such as traditional face-to-face education, computer-based educational systems and blended learning systems. Each of the different educational systems provides different data sources [5]. Using machine learning techniques, such as clustering, text mining, and classification techniques, these different types of data are analyzed to solve various educational problems. The taxonomy comprises thirteen tasks addressed by EDM systems, including predicting student performance, detecting undesirable student behaviors, profiling and grouping students, social network analysis, providing reports, creating alerts for stakeholders, planning and scheduling, creating courseware, developing concept maps, generating recommendation, adaptive systems, evaluation and scientific inquiry [6], [7].

D'Mello discussed the ubiquity and importance of emotion to learning [8]. The emotions may not always be consciously experienced, but they existed and influenced cognition nonetheless [9]. Language can express feelings very well, so text mining-based sentiment analysis techniques have great potential for analyzing the relationship between students' thoughts and learning experiences. Yang *et al.* applied sentiment analysis techniques on students' posts on MOOCs courses. They found that a negative correlation between the ratio of positive to negative terms and dropout across time [10]. Methods to automatically identify student confusion were developed from MOOCs posts [11]. This analysis method only uses unimodal data; MOOCs are now able to provide researchers with multimodal data, including students' behavioral data, textual data, audio, video, brainwave data, and more.

DataShop dataset was one of the first and biggest datasets that also provided a tool for intelligent tutoring systems [12]. While the student learned from the software, the student's actions and the tutor's responses were stored in a log database or file, which was imported into DataShop for storage and analysis. Graphical Interactive Student Monitoring Tool for Moodle (GISMO) is another popular public dataset and is a graphical interactive monitoring tool that provides useful visualization of students' activities in online courses to instructors. With GISMO instructors can examine various aspects of distance students, such as the attendance to courses, reading of materials and submission of assignments. Users of the popular learning management system Moodle may benefit from GISMO for their teaching activities [13]. Unimodal sentiment features and classifications (e.g., text, audio, and video) are used for sentiment discovery and analysis (SDA) [14]. The Multimodal Teaching and Learning Analytics (MUTLA) dataset was very well described and covered many academic subjects (i.e., Mathematics, English, Physics and Chemistry). User records at question level log of student responses, brainwave data and webcam data were collected [15]. The MUTLA dataset is the first rich multimodal dataset for EDM, but the MUTLA dataset is not open now. Cano *et al.* developed a multiview early warning system built with comprehensible Genetic Programming classification rules adapted to specifically target underrepresented and underperforming student populations. The system integrated many student information repositories using multi-view learning to improve the accuracy and timing of the predictions [16].

There are no open multimodal educational assessment datasets available. To address the lack of multimodal datasets, we collected multimodal data from several teaching management systems and MOOCs platforms. The data includes student behavior as well as students' course comments. The reason for choosing course comments instead of other data formats such as audio or video data is that the course commenting module exists in most MOOCs platforms. This makes data collection less expensive and our proposed multimodal data fusion model has a strong generalization capability.

Research on multimodal data fusion has focused more on the processing of text and images [17], however, the educational multimodal data fusion has not fully been exploited. To address the problem of heterogeneous data mining, students' behavior data and comment textual data are collected and manually aligned. Then a multimodal data fusion approach is designed to fuse structured students' behavior data and unstructured students' comment textual data into a unified semantic representation to predict student performance. Based on the dataset we collected, we conducted an empirical study. The study results show that the classification method can achieve better classification results in terms of RECALL, F1 and AUC.

In our study, to better elucidate our proposed research idea of multimodal data fusion for educational data mining, we design the following four research questions (RQs):

RQ1: Whether a multimodal dataset can be used to obtain a better classification model than a unimodal dataset?

RQ2: Can our proposed method outperform other data fusion methods when performing teaching effectiveness evaluation?

RQ3: Does our proposed model have strong a generalization ability?

RQ4: Can we perform interpretable analysis on our proposed deep multimodal data fusion model?

In summary, the contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to propose the use of student behavior data with course comments textual data to predict student performance.
- we are the first to propose an open dataset that includes student behavior data as well as course comments textual data.
- We are the first to propose a Transformer-based framework for creating deep multimodal data fusion algorithms with a uniform vector representation.

- Empirical results on real-world datasets show the effectiveness of our proposed method.

The rest of this paper is organized as follows. Section II introduces the background of educational data mining and multimodal data fusion. Section III describes our proposed method in detail, including the framework of deep teaching quality assessment based on multimodal data fusion, the Transformer-based semantic representation of course comment texts and deep multimodal data fusion algorithm. Section IV reports our experimental setup, including experimental subjects, performance evaluation measures, strategies for experimental comparison, and experimental design. Section V discusses the results of our experiments. Section VI analyzes the potential threats to the validity of our empirical results. Section VII concludes the paper with some future work.

II. BACKGROUND AND RELATED WORK

In this section, we mainly discuss the related studies on educational data mining, sentiment analysis, and multimodal data fusion.

A. EDUCATIONAL DATA MINING

Traditional educational institutions have accumulated a large amount of basic teaching data, such as basic information about students, through information transformation over the years. With the development of mobile Internet in recent years, web-based online educational systems have flourished exponentially, thus providing multiple data sources with different granularity levels [18]. With the emergence of more new teaching systems, such as MOOCs, massive amounts of data are constantly being collected. There is considerable value in these enormous teaching data. Many new research areas have been born for the new education system. Educational Data Mining is concerned with developing methods for exploring the unique types of data that come from educational environments [4]. Learning Analytics can be defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs [19]. Academic Analytics and Institutional Analytics are concerned with the collection, analysis, and visualization of academic program activities such as courses, degree programs research, the revenue of students' fees, course evaluation, resource allocation, and management to generate institutional insight [20]. Educational Data Science is defined as the use of data gathered from educational environments/settings for solving educational problems [18]. The different research areas share the same research interests using a data-driven approach to educational research and share the same goal of improving teaching and learning practices.

For intelligent tutoring systems, Markov Decision Process (MDP) framework was used to analyze and explore the application and effect of pedagogical strategies with EDM/LA techniques in terms of chi-squared, information gain, symmetrical uncertainty, information gain ratio and Weighted Information gain. The data source was collected in one per semester from the Spring of 2015 to the Fall of 2017 [21]. Chui *et al.* stated that improved conditional generative adversarial network based deep support vector machine (ICGAN-DSVM) algorithm was proposed to predict students' performance under supportive learning via school and family tutoring [22]. For learning management systems, Partial Least Squares Structural Equation Model (PLS-SEM) was used to analyze collaborative learning and to predict the team grade in teamwork groups. The data source was collected from a CS2 course [23]. For e-Learning Management Systems, an interpretable rule-based Genetic Programming classifier was used to predict student performance and students at risk as soon as possible to intervene early to facilitate student success in terms of Geometric mean, AUC, and Kappa. The student data was from the Virginia Commonwealth University [16]. In addition to analyzing computer-based educational systems from students' behavior data, sentiment analysis during online learning was also used to predict learning performance.

B. SENTIMENT ANALYSIS

Sentiment analysis (SA), also called Opinion Mining (OM) was the task of extracting and analyzing people's opinions, sentiments, attitudes, perceptions, etc. Sentiment analysis posed a powerful tool for researchers to extract and analyze public mood and views and finally make better decisions [24], [25]. SDA aims to automatically identify the underlying attitudes, sentiments, and subjectivity towards a certain entity such as learners and learning resources. Due to its enormous potential for smart education, SDA has been deemed a powerful technique for identifying and classifying sentiments from multimodal and multisource data over the whole process of education [14]. D'Mello discussed the ubiquity and importance of emotion to learning [8]. The emotions may not always be consciously experienced, but they existed and influenced cognition nonetheless [9]. Language can express feelings very well, so text mining-based sentiment analysis techniques have great potential to analyze the relationship between students' thoughts and learning experiences. Yang *et al.* applied sentiment analysis techniques to students' posts on their MOOCs courses. They found a negative correlation between the ratio of positive to negative terms and dropout across time [10]. Methods to automatically identify student confusion were developed from MOOCs posts [11].

C. MULTIMODAL DATA FUSION

Han *et al.* argued that there were many studies on unimodal sentiment features and classifications(e.g., text, audio and visual) [14]. Though they presented a novel SDA framework of multimodal fusions, together with the description of their crucial components, how to implement this multimodal framework had not been studied. The MUTLA dataset is the first rich multimodal dataset for EDM. Cano *et al.* developed a multiview early warning system built with comprehensible Genetic Programming classification rules adapted to target underrepresented and underperforming student populations [15]. The system integrated many student information repositories using multi-view learning to improve the accuracy and timing of the predictions [16]. For MOOCs courses, student behavior data can be obtained from the logs of the software system, and course comments can reflect the emotional state of the student learning process. The datasets for student behavior and course comments are easy to be collected and the cost of collecting these data is manageable compared to collecting brainwave data, video data, etc. Therefore, fusing student behavior data with course comments can better reflect the learning process of students and enable the prediction of student performance. Previous research in educational data mining has been conducted in a relatively isolated manner, either from student behavioural data or from the perspective of student sentiment analysis. It is difficult for such studies to comprehensively measure the behaviour of students during their online learning process. Especially with the popularity of MOOCS, more and more students are involved in the learning process and they express their attitudes towards the course by leaving comments. These student comments and student behaviour provide a good basis for our data modelling. we can predict student performance based on tabular and textual data.

III. OUR PROPOSED METHOD

In this section, we first briefly describe the framework of deep teaching quality assessment based on multimodal data fusion; then, the Transformer-based semantic representation of review texts and deep multimodal data fusion algorithm is proposed.

A. FRAMEWORK FOR DEEP EDUCATION QUALITY ASSESSMENT BASED ON MULTIMODAL FUSION DATA

Online education platforms, like MOOCs, provide a fast, interactive platform for educational data mining. From the MOOCs platform, students' learning process data can be collected, including both student behavior data and student interaction information, such as student comments on the learning course. The data can be extracted from relational databases at a low cost. We can intuitively feel that students who study hard will be more motivated to complete their assignments and will eventually achieve better performance. In addition, we can also just get the students' learning status from their course comment text. For example, students who are more optimistic about their course tend to have more positive attitudes toward learning, leading to better academic performance. The process of extracting data from a relational database is shown in Figure 1. Based on the Transformer architecture's powerful learning capability of natural language, we have the potential to learn more information about students' learning status from course comments, which will ultimately enhance deeper mining of student learning data.

The student learning process data of different modalities contain rich user information, and data mining can be performed for the student learning process data of different modalities to build a student teaching quality assessment model. The framework for deep education quality assessment is shown in Figure 2, which is based on the semantic vector representation of students' comment text, as well as students' behavior data. In the deep teaching quality assessment framework, the problem of predicting student performance is formalized as a binary classification problem, and the model classifies the results as excellent learning effect or average learning effect. The feature vector classification function is defined as:

$$y' = \mathrm{argmax}_{c\in\{0,1\}} f_\theta(x) \tag{1}$$

In Equation 1, x represents the input student learning status data, including student behavior data, such as MOOC learning progress, learning progression for objective practice questions, etc., and also includes the students' course comments, for example, the student's comment, "The course is rather obscure and covers a lot of underlying principles.". Student behavior data and course comments are persistently stored in a relational database from online education platforms, such as MOOC and SPOC Academy, as well as from third-party open data interfaces, such as Golden Classroom. $f_\theta(.)$ denotes the classifier obtained by historical training data of student learning, such as random forest, etc. The training data of the model is done by aligning multiple databases, and the excellent learning effect is labeled as 1, and the average learning effect is labeled as 0. For the training dataset D_{tr}, a dataset containing N training samples is defined $D_{tr} = \{x_n, y_n\}_{n=1}^{N}$, the samples are labeled $y_n \in \{0, 1\}$,and the training samples $x_n = \left(x_n^1, x_n^2, x_n^3, x_n^4, x_n^5, x_n^6, x_n^7\right)$, x_n^1 to x_n^6 denote the behavioral characteristics of student learning, including learning progress (LP), learning progression for objective practice questions (LPO), learning progression for subjective practice question (LPS), in-class discussion participation (DP), number of posts and number of replies respectively. The definitions of each behavioral characteristic are as follows.

$$\mathrm{LP} = \frac{\text{Number of studied chapters}}{\text{Total number of course chapters}} \tag{2}$$

$$\mathrm{LPO} = \frac{\text{Number of completed objective questions}}{\text{Total number of objective questions}} \tag{3}$$

$$\mathrm{LPS} = \frac{\text{Number of completed subjective questions}}{\text{Total number of subjective questions}} \tag{4}$$

$$\mathrm{DP} = \frac{\text{Number of submitted class exercises}}{\text{Total number of class exercises}} \tag{5}$$

The number of posts and the number of replies refer to the number of posts made by students in the forum of the MOOC platform. The above data features are collected from the MOOC platform, which are exported after students finish a course on the MOOC platform. x_n^7 represents the one-dimensional feature vector of students' course comments, as shown in Figure 2, which is computed from a deep semantic vector learning model based on Transformer. x_n is

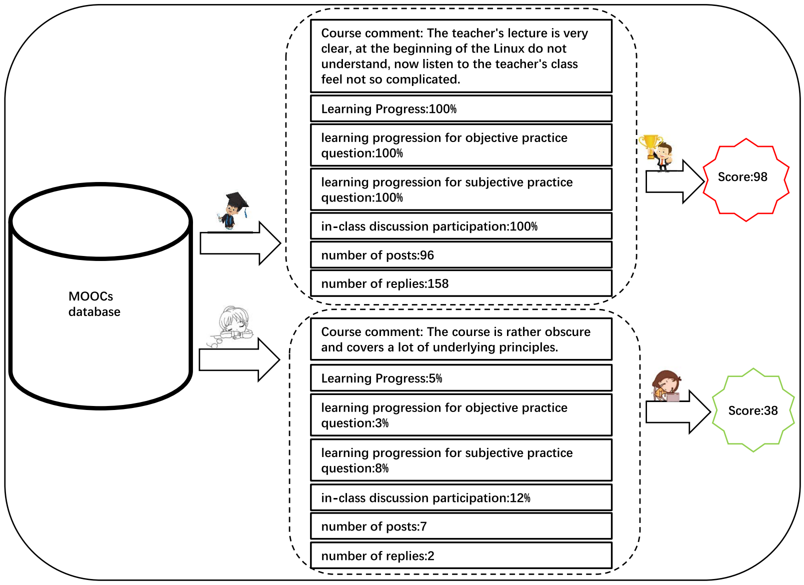


FIGURE 1. The process of extracting data from a relational database.

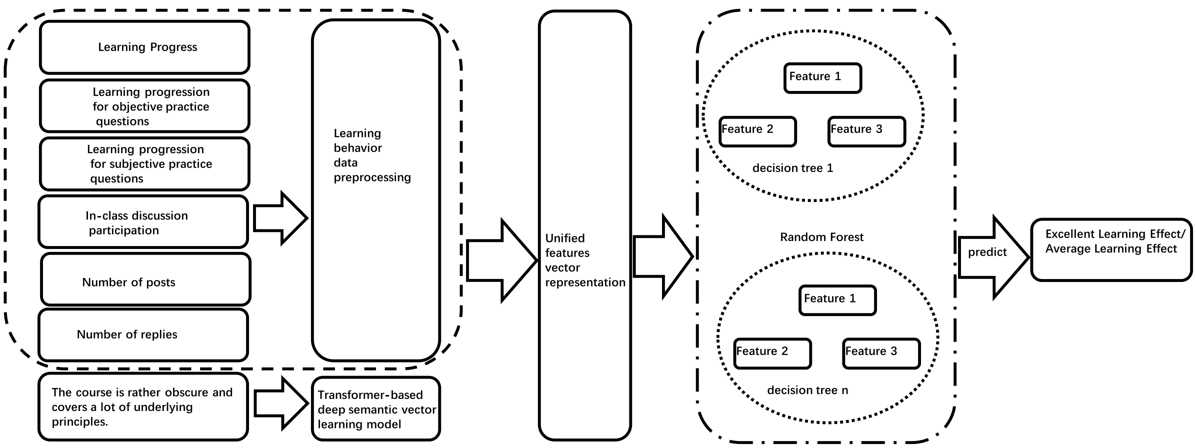


FIGURE 2. An overview of our study.

represented as a uniform feature vector of student learning state data. Multiple decision trees are constructed to form a random forest-based on the training data D_{tr} and ensemble learning are used in the random forest. x_n^7 differs from other features in that its conditional entropy must be calculated considering the domain feature migration of the Transformer network, and the parameters of the Transformer network are determined based on the training data of the review text, and its specific calculation formula is as Eq. 6.

$$H(D \mid A) = w_\theta \left(\sum_{n=1}^{N} \frac{|D_n|}{|D|} H\left(D_n\right) \right) \tag{6}$$

w_θ represents the Transformer network that determines the optimal network parameters, $H(D \mid A)$ denotes the empirical conditional entropy in the case of condition A of the selected information entropy calculation, $|D_n|$ indicates the number of samples for a given classification for a selected characteristic, $\frac{|D_n|}{|D|}$ indicates the probability of a classification for a selected feature, $H\left(D_n\right)$ denotes the empirical information entropy of D.

B. THE TRANSFORMER-BASED SEMANTIC REPRESENTATION OF REVIEW TEXTS

The Transformer architecture has gained wide application in natural language processing. The pre-trained BERT models can achieve better classification performance after fine-tuning domain-specific data, and its classification is done by computing cross-entropy loss functions on feature vectors by a linear classifier [26]. The attention mechanism and the feature vector representation of text provide a unified representation for the fusion of multimodal data, such as

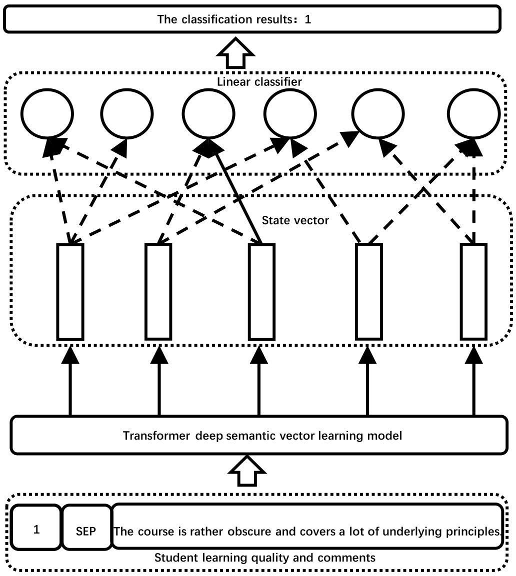


FIGURE 3. Transformer-based learning method for 1D text feature vector representation.

text and images [17]. The multidimensional feature vector representation of text is unstructured data, and fusing the table-based data from the teaching process with the feature vector directly or calculating the attention between different features does not fully use the features of the table-based data [27], and conversely, integrating the multidimensional feature vector of BERT fine-tuning into the table-based data may bring the problem of feature redundancy. To construct the teaching quality assessment model, a Transformer-based 1D text feature vector representation learning method is designed, and this deep semantic feature learning process is shown in Figure 3. We give an example of the learning process for eigenvectors of student comments. For example, the student comment "The course is rather obscure and covers a lot of underlying principles". The student grade of 60 is converted to label 1. The student comment is fed as input to a pre-trained model for fine-tuning. A linear classifier is used at the feature vector layer to classify the output features on a multilayer neural network, and a loss function is calculated by comparing it with the classification label 1.

The course review texts of students and the results of the teaching quality assessment are taken as input, and the course review texts are fine-tuned using the Transformer-based model BERT. The deep semantic connection between the input texts and the teaching quality assessment is established based on the attention mechanism. In the input layer, the word embedding vector representation of the review text is performed. The segment embedding representation of each text, as well as the location embedding vector are obtained, and the summation is the vector input of each review text. The key to learning the semantic vector of course review texts is to use the multiple attention mechanism to obtain the connection between course review texts and teaching quality assessment results, which is done by fine-tuning the course review texts using BERT. The fine-tuned model outputs a state vector, which is used as the input to a linear classifier for learning. This linear classifier is defined as shown in Eq. 7.

$$X_{\text{output}} = \text{Linear}\left(\text{ReLU}\left(\text{Linear}\left(X_{\text{attention}}\right)\right)\right) \tag{7}$$

$X_{\text{attention}}$ represents the state vector output of the pre-trained model, and X_{output} is the output result, which represents the classification result of the comment text, with the value of 0 or 1. The loss function of the current training sample is calculated based on the outcome of the binary classification, and the cross-entropy loss function used in the calculation is shown in Eq. 8.

$$L = \frac{-1}{2}\left((1-\alpha)y_i \log(p_i) + \alpha\,(1-y_i)\log(1-p_i)\right) \tag{8}$$

L denotes the calculated loss function value, y_i denotes the actual probability that the sample is i, p_i denotes the predicted probability obtained based on the training of the fine-tuned BERT model on the historical dataset, and α denotes the proportion of classes with actual probability p_i on the training dataset over the total dataset. α parameter is used to address the class imbalance problem existing in the training dataset [28]. Fine-tuning of this teaching quality assessment model was completed after recording the best classification model on the validation dataset. The training dataset is reintroduced into the final Transformer model and the semantic representation of the review text for this historical data is computed by forwarding computation.

C. DEEP MULTIMODAL DATA FUSION ALGORITHM

As shown in Figure 2, the deep multimodal fused data-based teaching quality assessment framework uses a random forest classifier to classify a uniform feature vector and create multiple decision trees to vote to predict student learning effectiveness. This unified feature vector is the Transformer-based deep multimodal data fusion representation, which includes both behavioral data during student learning and Transformer-based comment text semantic vectors. The deep multimodal data fusion process is shown in Algorithm 1.

As shown in Algorithm 1, the algorithm can effectively use the tabular data of students' learning behavior and meanwhile, embed the one-dimensional feature vector of comment text into the tabular data. So random forest can fully use the information entropy of unified features vector representation to build students' learning quality assessment model. The algorithm can effectively integrate with the traditional online teaching platform to extract students' behavior from the relational database; at the same time, the algorithm introduces students' interaction behavior of comment text, enriching the description of student learning status and describing the student learning process from more dimensions.

Algorithm 1 Deep Multimodal Data Fusion Algorithm

Input :
training set $D_{tr} = \{x_n, y_n\}_{n=1}^{N}$;
pre-trained model BERT;

Output:
unified features vector representation R^x;

for *data in* D_{tr} **do**
  Feed forward x_n^7 in BERT, compute loss value and back propagation;
  Record the neural network parameters and obtain the domain representation of the text;
end
for *data in* D_{tr} **do**
  Freeze deep neural networks and perform forward pass;
  Obtain a one-dimensional semantic vector of comment text v_{text} ;
  Concatenate, $R_i^x = \left(x_i^1, x_i^2, x_i^3, x_i^4, x_i^5, x_i^6, v_{\text{text}}\right)$;
end
Use R^x to train random forest classifier RF_{quality} ;

TABLE 1. The teaching quality assessment dataset.

Label	Data Description	Samples	Ratio
0	Samples with student grades≥ 85	4420	70.30%
1	Samples with student grades <85	1870	29.70%

IV. EXPERIMENTAL SETUP

In this section, we introduce the experiment setup, including experimental subjects, performance evaluation metrics, multimodal data fusion methods and experimental design.

A. EXPERIMENTAL SUBJECTS

To compare the data fusion methods, we collected one dataset for predicting student performance and used one publicly available dataset to evaluate the generalizability of our proposed method.

The first dataset we collected comes from the MOOCs platform we are using. The courses are intended for college students. The collected data comes from three teaching systems, including a MOOCs platform, the student course evaluation system and academic management system. Learning progress, learning progression for objective practice question, learning progression for subjective practice question, in-class discussion participation, number of posts and number of replies were obtained from a MOOCs platform; the students' comments were obtained from the student course evaluation system. Students' course grades were obtained from academic management system. A brief description of the first teaching quality assessment dataset is described in Table 1.

The second dataset is Women's E-Commerce Clothing Reviews dataset, collected by Nick Brooks in 2018. This dataset is used to evaluate the generalization ability of our proposed method on a publicly available dataset and has been used to perform binary classification [27]. The source of the reviews is anonymous. Data examples consist of a review, a rating, the clothing category of the product etc.

TABLE 2. Confusion matrix for predicting student performance.

	Predicted Average	Predicted Excellent
Actually Average	True Positive (TP)	False Negative(FN)
Actually Excellent	False Positive (FP)	True Negative(TN)

B. PERFORMANCE EVALUATION METRICS

There is class imbalance in the teaching quality assessment dataset. We consider three performance metrics: recall, F1-measure (F1) and the area under the receiver operating characteristic curve (AUC). The confusion matrix for the teaching quality assessment dataset is shown in Table 2, *TP* (true positive) indicates that the sample with average learning effect is correctly predicted as average, *FN* (false negative) indicates that the sample with average learning effect is incorrectly predicted as excellent, *FP* (false positive) indicates that the sample with excellent learning effect is incorrectly predicted as average, and *TN* (true negative) indicates that the sample with excellent learning effect is correctly predicted as excellent.

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

$$F1 = \frac{2 \times (precision \times recall)}{precision + recall} \tag{12}$$

The AUC is calculated as the area formed by the Receiver Operating Characteristic (ROC) curve and the coordinate axis, with the maximum value not exceeding 1. The larger the AUC value, the better the classification effect. The *TPR* indicates the percentage of samples that are correctly predicted as average learning effect among all samples that are actually average learning effect, and its value is equal to recall; the *FPR* indicates the percentage of samples that are incorrectly predicted as average learning effect among all samples that are actually excellent learning effect.

To statistically evaluate the detailed results, we first employ the Friedman test to determine whether there are statistically significant differences among compared methods. If there is a statistically significant difference, the post-hoc Nemenyi test is applied to compare the difference.

When the null hypothesis is rejected, the average rank should be calculated and compared with the critical distance (CD).

$$CD = q_a \times \sqrt{\frac{k \times (k+1)}{6N}} \tag{13}$$

k represents different algorithms, and N represents all training datasets. q_a is obtained by looking up the table depending on the different parameters. Therefore, the result of CD can be computed according to the Eq. 13. In addition, to evaluate the degree of difference among the compared methods in terms of recall, F1 and AUC, we apply Cohen's d to measure the effect size [29], [30], [31].

$$Cohen's\ d = \frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}} \tag{14}$$

where M_1 and M_2 represent the mean of the statistic, and σ represents the standard deviation of the statistic. If $d \in \{0, 0.2\}$, this indicates the effect size is negligible. If $d \in \{0.2, 0.5\}$, this indicates the effect size is negligible. If $d \in \{0.5, 0.8\}$, this indicates the effect size is medium. If $d \in \{0.8, 1\}$, this indicates the effect size is large.

C. MULTIMODAL DATA FUSION METHODS

To evaluate our proposed teaching quality prediction model (RfBERT) in a comprehensive manner, we chose the following data fusion methods as baseline methods for comparison.

text_only: As shown in Eq. 8, only the review text is used as the input of the teaching quality assessment model, and the cross-information entropy is used as the loss function to build the Transformer-based text classification model.

tabular_only: Using $x_n = \left(x_n^1, x_n^2, x_n^3, x_n^4, x_n^5, x_n^6\right)$ as input samples, we build a teaching quality assessment model based on random forest.

concat: As shown in Figure 2, the uniform feature vector is used as the input sample to build a teaching quality assessment model with a linear classifier.

MLPconcat: Implementation of the data fusion method proposed by Gu et al [27]. This method separate MLPs on numerical feats then concatenation of transformer output, with processed numerical feats before the final classifier layer(s).

MAG: Implementation of the data fusion method based on the attention mechanism proposed by Rahman1 *et al.* [32]. In the output layer, this method used gated summation of transformer outputs, numerical feats, and categorical feats before the final classifier layer(s).

D. EXPERIMENTAL DESIGN

The experiments run on Windows OS, and the running hardware environment is Intel Core i7-10700K CPU with 64G RAM. The fine-tuning of BERT is completed on NVIDIA GeForce RTX 2070 GPU. The deep neural network library used in the experiments is Pytorch 1.8 stable version and the open-source huggingface library is used to implement BERT.

The pre-trained model BERT used in the experiments is bert-base-uncased, which has 12 layers, 12 head attention, word embedding dimension of 768 and network parameters of 110 M. Sentences exceeding a specific length are truncated, and zero-fill operations are performed for sentences that do not satisfy the length. The open source sklearn framework is used, and the random forest classifier uses the default hyperparameters, where the number of classification subtrees is 10. The ratio of training dataset, validation dataset and test dataset was 8:1:1 during the experiment, and the number of training repetitions was 10, with random stratified sampling each time to maintain the consistency of data distribution. During the validation process, the early stop method was used to terminate the neural network training process to prevent overfitting. In the fine-tuning of the Transformer model, a hyperparameter α is introduced to solve the class imbalance problem in the training data set, which reduces the impact of the majority class on the imbalanced data distribution by penalizing the loss value of the majority class.

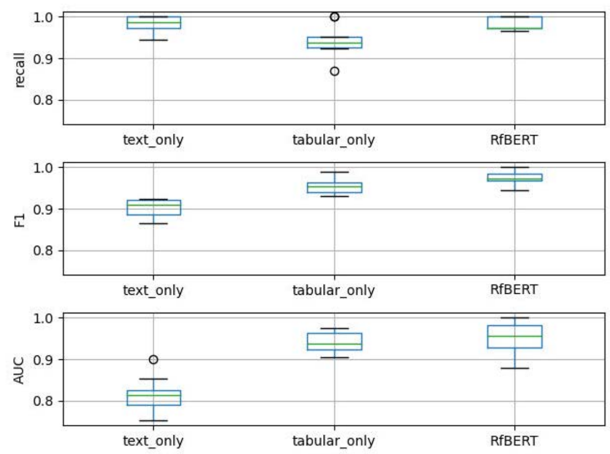


FIGURE 4. The comparison results via box plot.

V. EXPERIMENTAL RESULTS

In this section, we report experimental results for the four RQs.

A. RESULT ANALYSIS FOR RQ1

RQ1: Whether a multimodal dataset can be used to obtain a better classification model than a unimodal dataset?

Motivation: To verify whether better performance in predicting student performance can be obtained by fusing multimodal data, we compare our proposed method with classifiers that employ unimodal data. The text_only fusion method is preformed to compare the tabular data and the tabular_only fusion method is performed to compare the textual data.

To answer this RQ, we conduct the experiments on the collected dataset. The comparison results via box plot are shown in Figure 4. From these figures, we can observe that our proposed method RfBERT achieves best performance in terms of recall, F1 and AUC. All three different data fusion methods obtained high recall; the AUC value of the text data fusion method was only 0.8153 and the method also had the lowest performance on the F1 metric, which could indicate that student performance could not be fully predicted using only the text of student course comments. The tabular_only fusion method achieves sub-optimal performance in F1 and AUC metrics; our proposed method achieves the best performance in all three metrics. This means that our proposed method can fully fuse two different kinds of data by learning feature vectors from text and then achieves the best performance.

To compare the performance of different data fusion methods from a statistical point of view, the non-parametric Friedman test at a confidence level of 95% is used to conduct a statistical analysis of the results. We find that the calculated value is smaller than the critical value for a 0.05 significance level. To reveal the differences between different data fusion methods, we further adopt a post hoc statistical analysis method. In this experiment, k means three different algorithms, and $N = 10$ means that the collected dataset was randomly sampled 10 times. Finally, Cohen's d effect size is 0.75 between RfBERT and text_only in terms of AUC and this indicates the effect size is medium; Cohen's d effect size is 0.82 between RfBERT and tabular_only in terms of AUC and this indicates the effect size is large.

Summary for RQ1: From the box plot as well as the statistical results, our proposed method can fully fuse two different kinds of data by learning feature vectors from text and then achieves the best performance. Course comment texts should be considered when creating student academic assessment models.

B. RESULT ANALYSIS FOR RQ2

RQ2: Can our proposed method outperform other data fusion methods when performing teaching effectiveness evaluation?

Motivation: Based on Gu *et al.* research [27], for multimodal data fusion, there are three main baselines,including concat, NLPconcat, MAG. We need to verify whether our proposed method can outperform the three baseline methods.

According to the box plot of Figure 5, all four data fusion methods obtained high recall, while our proposed data fusion method obtained the highest recall. In terms of the AUC and F1 metrics, the MAG data fusion method obtained the lowest performance and the concat data fusion method obtained the sub-optimal performance. Compared with concat, MLPconcat incorporates MLP on the tabular data output layer, which may be the reason for its performance degradation. We can clearly see that our proposed method obtains the best performance and has a high stability.

The non-parametric Friedman test at a confidence level of 95% is used to conduct a statistical analysis of the results. We find that the calculated value is smaller than the critical value for a 0.05 significance level. The post hoc statistical analysis method was adopted. In this experiment, k means four different algorithms, and $N = 10$ means that the collected dataset was randomly sampled 10 times. Finally, Cohen's d effect size is 0.68 and this indicates the effect size is medium.

Summary for RQ2: Our proposed method achieves the best classification performance compared to the base methods. This implies that the uniform feature vector representation learned by our proposed method can indeed improve the classifier's performance.

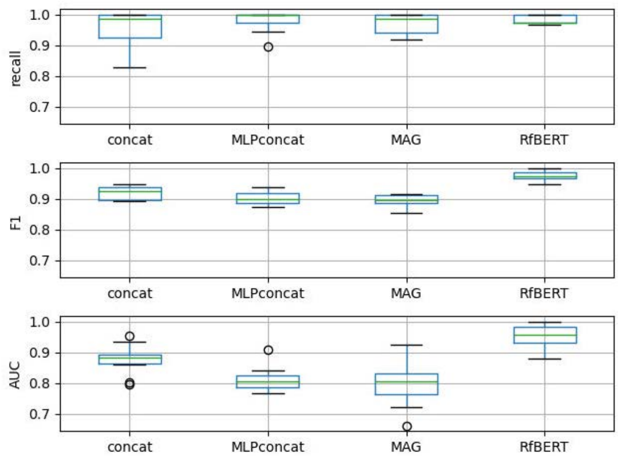


FIGURE 5. The box plot on different data fusion methods.

TABLE 3. Performance comparison of data fusion methods on open dataset.

data fusion method	F1	AUC
concat	0.969909	0.97923
MLPconcat	0.959	0.9622
MAG	0.961	0.963
RfBERT	0.991	0.985

C. RESULT ANALYSIS FOR RQ3

RQ3: Does our proposed model have strong generalization ability?

Motivation: Although our method achieves the best classification performance on the dataset we collected, to validate the generalization ability of our proposed method, we compare multiple feature fusion methods on an open dataset.

We conducted experiments on the open clothes review dataset to evaluate feature fusion methods including concat, MLPconcat, MAG, RfBERT. We performed a ten fold cross-validation and obtained the mean F1 and AUC. The results are shown in Table 3.

Our proposed method RfBERT obtains the best classification performance on F1 and AUC metrics. On the F1 metric, our method improves 3.34% over the method MLPconcat, which achieves the worst performance. Meanwhile, on the AUC metric, our method improves 2.37% over the method MLPconcat, which achieves the worst performance. As similar to the results of the RQ2 experiment, concat obtained suboptimal performance on both F1 and AUC metrics. The performance of MAG is slightly stronger than MLPconcat and lower than our proposed method. This conclusion remains largely consistent with RQ2. This implies that for multimodal tabular and textual data, unified features vector representation can effectively improve the classification performance.

Summary for RQ3: Our proposed method has a strong generalization capability. The classification performance can

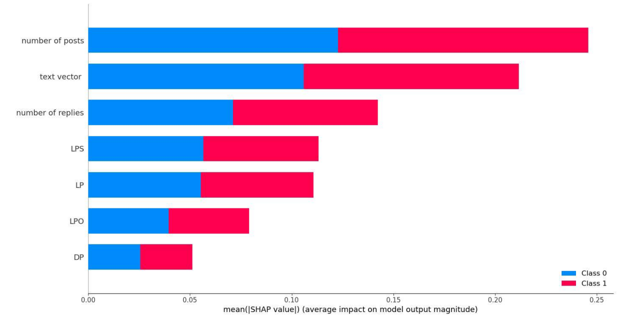


FIGURE 6. The importance of each feature using SHAP method.

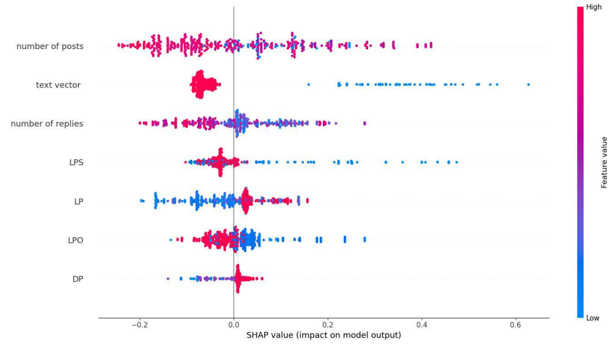


FIGURE 7. The contribution of different features to the prediction result of "average learning".

be improved by using a unified features vector representation for multimodal tabular and textual data.

D. RESULT ANALYSIS FOR RQ4

RQ4: Can we perform interpretable analysis on our proposed deep multimodal data fusion model?

Motivation: Based on RQ1, RQ2 and RQ3, we can build a prediction model with strong generalization ability and higher performance from the historical dataset. The model contains multimodal data. To analyze students' academic performance in a timely manner and intervene accordingly, it is necessary to conduct an interpretable analysis of our model.

To evaluate the contribution of the seven features in the unified feature vector to the random forest classifier, we introduce the SHAP method. The calculation procedure is shown in Eq. 15. Suppose the i sample is x_i, the j feature of the i sample is $x_{i,j}$, the predicted value of the model for the i sample is y_i, and the baseline (usually the mean of the target variable for all samples) of the whole model is y_{base}.

$$y_i = y_{\text{base}} + f\left(x_{i,1}\right) + f\left(x_{i,2}\right) + \cdots + f\left(x_{i,7}\right) \tag{15}$$

We implemented the SHAP method on the random forest classifier and calculated the importance of each feature as shown in Figure 6. We see that the feature "number of posts" is the largest contributor to the prediction result, followed by the text vector we learned from the BERT. This also shows that the course comment texts we used can improve the performance of the classifier and make an important contribution to the classification task.

Further we can observe the contribution of different features to the prediction result of "average learning". The result is shown in Figure 7. A total of seven features influence the classification results of the model. When the classification result of the model is "average learning", different features contribute differently. The first feature and the sixth feature have a relatively even effect on the current classification result; the second feature has a negative effect on the current classification result most of the time. The second feature has a negative effect on the current classification result, indicating a negative correlation between the student's course comments and the current classification result. The third feature has an average impact on the classification results. The fourth feature also shows a negative correlation with the classification results. The fifth and seventh characteristics show a positive relationship with the classification results. We can see that the learned text feature vector plays an important role for the category "average learning". This is consistent with the distribution of our data.

Summary for RQ4: Based on the results of the interpretability analysis, we see that the unified feature vector fused with the text vector can indeed play a key role in model classification.

VI. THREATS TO VALIDITY

In this section, we mainly discuss potential threats to the validity of our study.

A. THREATS TO CONSTRUCT VALIDITY

To evaluate our proposed approach, we collected data from multiple instructional management systems and built experimental datasets by alignment. However, the size of these data is currently small, and the dataset will need to be continuously expanded later. When testing the generalization capability, the test was conducted on only one open dataset. As more datasets are shared, there is a need to validate on more datasets.

B. THREATS TO INTERNAL VALIDITY

We use several open source software in our experiments, such as huggingface, sklearn, etc. These open source software provide default hyperparameter settings, such as pre-trained model BERT, etc. Although we fine-tuned the deep neural network by validation dataset, there are still more hyperparameters with default values. In addition, the machine learning classifiers we used, such as random forest, also used the default hyperparameter settings.

C. THREATS TO EXTERNAL VALIDITY

External validity is the degree to which the research results can be generalized to the population under study and other research settings. There are no commercial datasets available for testing yet, and we need to keep an eye on developments based on multimodal tabular and textual data fusion.

VII. CONCLUSION AND FUTURE WORK

With the emergence of more new teaching systems, such as MOOCs, massive amounts of data are constantly being collected. This massive amount of data is a vast gold mine. However, the multimodal data including both student behavior data and student course comments textual data, is not processed to discover models and paradigms which can be useful for school management. All these state data during the learning process can reflect the effectiveness of student learning. There is no multimodal dataset with tabular data and textual data yet. So we first collected an open dataset that included student behavior data as well as course comments textual data. We fused student behavior data with course comments textual data to predict student performance. Then a Transformer-based framework for creating deep multimodal data fusion algorithms with a uniform vector representation was proposed. The empirical results of the collected dataset show the effectiveness of our proposed method in terms of recall, F1 and AUC. The empirical research indicates that: (1)our proposed method can fully fuse two different kinds of data by learning feature vectors from text and then achieves the best performance. Course comment texts should be considered when creating student academic assessment models; (2)Our proposed method achieves the best classification performance compared to the base methods. This implies that the uniform feature vector representation learned by our proposed method can indeed improve the classifier's performance.

Further, we validated our approach on an open clothing dataset. The results of the empirical study showed that our proposed method had a strong generalization capability. Moreover, we performed interpretability analysis using SHAP method and found that text features had more important influence on the classification model. This further illustrated that fusing text features can improve the performance of classification models.

In the future, we will continue to expand our dataset and apply our proposed method to other domains to validate its generalization capability continuously. In addition, we will also continue our in-depth research on the representation of unified feature vectors based on natural language processing techniques.We will work on additional ways to fuse data to improve the classification performance of student learning classification models.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data science in massive open online courses," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 7, no. 1, p. e1187, Jan. 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1187 and https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1187 and https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1187

[2] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*. Springer, Jan. 2014, pp. 61–75. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4614-3305-7_4

[3] M. I. Baig, L. Shuib, and E. Yadegaridehkordi, "Big data in education: A state of the art, limitations, and future research directions," *Int. J. Educ. Technol. Higher Educ.*, vol. 17, no. 1, pp. 1–23, Dec. 2020. [Online]. Available: https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-020-00223-0

[4] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Educ. Inf. Technol.*, vol. 23, no. 1, pp. 537–553, Jul. 2017. [Online]. Available: https://link.springer.com/article/10.1007/s10639-017-9616-z

[5] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Exp. Syst. Appl.*, vol. 33, no. 1, pp. 135–146, Jul. 2007.

[6] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A systematic review of deep learning approaches to educational data mining," *Complexity*, vol. 2019, May 2019, Art. no. 1306039.

[7] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, "Performance analysis of the impact of technical skills on employability," *Int. J. Performability Eng.*, vol. 17, no. 4, p. 371, Apr. 2021. [Online]. Available: http://www.ijpe-online.com/EN/10.23940/ijpe.21.04.p5.371378

[8] C. Lang, G. Siemens, A. Wise, and D. Gasevic. (2017). *Handbook of Learning Analytics*. [Online]. Available: https://www.academia.edu/download/56326181/hla17.pdf

[9] A. öhman and J. J. Soares, "Unconscious anxiety': Phobic responses to masked stimuli," *J. Abnormal Psychol.*, vol. 103, no. 2, pp. 231–240, 1994.

[10] M. Wen, D. Yang, and C. P. Rosé. *Sentiment Analysis in MOOC Discussion Forums: What Does It Tell Us?* Citeseer. Accessed: Apr. 2022. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.660.5804&rep=rep1&type=

[11] D. Yang, M. Wen, I. Howley, R. Kraut, and C. Rosé, "Exploring the effect of confusion in discussion forums of massive open online courses," in *Proc. 2nd ACM Conf. Learn. Scale*, Mar. 2015, pp. 121–130.

[12] (2010). *A Data Repository for the EDM Community*. [Online]. Available: https://www.researchgate.net/publication/254199600_A_Data_Repository_for_the_EDM_Community

[13] *Graphical Interactive Student Monitoring Tool for Moodle*. Accessed: Apr. 2022. [Online]. Available: http://gismo.sourceforge.net/index.html

[14] Z. Han, J. Wu, C. Huang, Q. Huang, and M. Zhao, "A review on sentiment discovery and analysis of educational big-data," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 10, no. 1, p. e1328, Jan. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1328 and https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1328 and https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1328

[15] F. Xu, L. Wu, K. P. Thai, C. Hsu, W. Wang, and R. Tong, "MUTLA: A large-scale dataset for multimodal teaching and learning analytics," Oct. 2019, *arxiv:1910.06078v1*.

[16] A. Cano and J. D. Leonard, "Interpretable multiview early warning system adapted to underrepresented populations," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 198–211, Apr. 2019.

[17] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," 2019, *arXiv:1909.02950*.

[18] C. Romero and S. Ventura, "Educational data science in massive open online courses," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 7, no. 1, p. e1187, Jan. 2017.

[19] C. Lang, G. Siemens, A. Wise, D. Gašević, and A. Research. *Handbook of Learning Analytics Society for Learning*. Accessed: Apr. 2022. [Online]. Available: https://www.solarresearch.com

[20] J. Campbell, P. DeBlois, D. Oblinger. (2007). *Academic Analytics: A New Tool for a New Era*. [Online]. Available: https://er.educause.edu/articles/2007/7/academic-analytics-a-new-tool-for-a-new-era

[21] *Exploring Induced Pedagogical Strategies Through a Markov Decision Process Framework: Lessons Learned*. Accessed: Apr. 2022. [Online]. Available: https://par.nsf.gov/biblio/10105557

[22] K. T. Chui, R. W. Liu, M. Zhao, and P. O. de Pablos, "Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine," *IEEE Access*, vol. 8, pp. 86745–86752, 2020.

[23] Z. Li and S. Edwards. (2018). *Applying Recent-Performance Factors Analysis to Explore Student Effort Invested in Programming Assignments*. [Online]. Available: https://search.proquest.com/openview/1344abae126cd4240dfdce3764087786/1?pq-origsite=gscholar&cbl=1976352

[24] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134.

[25] A. G. Etemad, A. I. Abidi, and M. Chhabra, "Fine-tuned T5 for abstractive summarization," *Int. J. Performability Eng.*, vol. 17, no. 10, pp. 900–906, Oct. 2021. [Online]. Available: http://www.ijpe-online.com/EN/10.23940/ijpe.21.10.p8.900906

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[27] K. Gu and A. Budhkar, "A package for learning on tabular and text data with transformers," in *Proc. 3rd Workshop Multimodal Artif. Intell.*, 2021, pp. 69–73. [Online]. Available: https://aclanthology.org/2021.maiworkshop-1.10

[28] L. Fang, Q. Yubin, C. Xiang, L. Long, and Y. Fan, "A sentiment analysis method based on class imbalance learning," *J. Jilin Univ. Sci. Ed.*, vol. 59, no. 4, pp. 929–935, 2021. [Online]. Available: http://xuebao.jlu.edu.cn/lxb/CN/abstract/abstract4404.shtml

[29] Y. Qu, X. Chen, F. Li, F. Yang, J. Ji, and L. Li, "Empirical evaluation on the impact of class overlap for EEG-based early epileptic seizure detection," *IEEE Access*, vol. 8, pp. 180328–180340, 2020.

[30] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for T-tests and ANOVAs," *Frontiers Psychol.*, vol. 4, p. 863, Nov. 2013. [Online]. Available: /pmc/articles/PMC3840331//pmc/articles/PMC3840331/?report=abstract and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3840331/

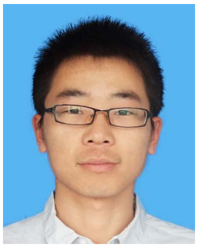
[31] S. S. Sawilowsky, "New effect size rules of thumb," *J. Modern Appl. Stat. Methods*, vol. 8, no. 2, p. 26, Nov. 2009. [Online]. Available: https://digitalcommons.wayne.edu/jmasm/vol8/iss2/26

[32] W. Rahman, M. Kamrul Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," 2019, *arXiv:1908.05787*.

YUBIN QU was born in Nanyang, China, in 1981. He received the B.S. and M.S. degrees in computer science and technology from Henan Polytechnic University, China, in 2004 and 2008, respectively. Since 2009, he has been a Lecturer with the Information Engineering Institute, Jiangsu College of Engineering and Technology. He is the author of more than ten articles. His research interests include software maintenance, software testing, and machine learning.

FANG LI was born in Baoji, China, in 1982. She received the M.S. degree in computer science and technology from Henan Polytechnic University, China, in 2011. Since 2014, she has been a Lecturer with the Jiangsu College of Engineering and Technology. Her research interests include network ideological and political education and computer application.

LONG LI (Member, IEEE) received the Ph.D. degree from the Guilin University of Electronic Technology, Guilin, China, in 2018. He is currently a Lecturer with the School of Computer Science and Information Security, Guilin University of Electronic Technology. His research interests include cryptographic protocols, privacy-preserving technologies in big data, and the IoT.



XIANZHEN DOU was born in Xuzhou, China, in 1987. He received the M.S. degree from the School of Electronics and Information, Nantong University, China, in 2013. Since 2019, he has been a Lecturer with the Information Engineering Institute, Jiangsu College of Engineering and Technology. His research interests include software engineering and machine learning.

HONGMEI WANG was born in Lianyuan, China, in 1981. She received the B.S. and M.S. degrees in computer science and technology from Henan Polytechnic University, China, in 2005 and 2008, respectively. She is currently pursuing the Ph.D. degree with the Nanjing University of Posts and Telecommunications. She was a Visiting Scholar at the University of Hong Kong, from 2018 to 2019. Since 2008, she has been with the Jiangsu University of Science and Technology. Her research interests include information security, machine learning, and artificial intelligence.