JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 02/06 pp12−29
Volume 4, Number 1, March 2023
DOI: 10.23919/JSC.2023.0007

Prediction of Academic Performance of Students in Online Live Classroom Interactions —An Analysis Using Natural Language Processing and Deep Learning Methods

Yuanyi Zhen, Jar-Der Luo*, and Hui Chen

Abstract: Prior studies have shown the importance of classroom dialogue in academic performance, through which knowledge construction and social interaction among students take place. However, most of them were based on small scale or qualitative data, and few has explored the availability and potential of big data collected from online classrooms. To address this issue, this paper analyzes dialogues in live classrooms of a large online learning platform in China based on natural language processing techniques. The features of interactive types and emotional expression are extracted from classroom dialogues. We then develop neural network models based on these features to predict high- and low-academic performing students, and employ interpretable AI (artificial intelligence) techniques to determine the most important predictors in the prediction models. In both STEM (science, technology, engineering, mathematics) and non-STEM courses, it is found that high-performing students consistently exhibit more positive emotion, cognition and off-topic dialogues in all stages of the lesson than low-performing students. However, while the metacognitive dialogue illustrates its importance in non-STEM courses, this effect cannot be found in STEM courses. While high-performing students in non-STEM courses show negative emotion in the last stage of lessons, STEM students show positive emotion.

Key words: academic performance prediction; live classroom dialogue; emotional expression; interactive type; natural language processing; deep learning

1 Introduction

While various studies have examined the relation between in-class interactions and students' academic performance, there has few works which investigate this issue in online education by big-data analysis. We fill this gap by collecting academic performance and in-class interaction data from 89 694 STEM (science, technology, engineering, mathematics) course students and 32 630 non-STEM course students in a large Chinese online education platform. This study employs two classroom dialogue classification models to extract interaction features from live classroom dialogues, and descriptive statistics and deep learning models with interpretable AI (artificial intelligence) to determine the most important predictors of students' performance. Through further big data analyses, we investigate the similarities and differences between STEM and non-STEM courses.

With the increasing penetration of technologies such as live streaming and computer-assisted instruction, an increasing number of students can take classes and receive more effective instruction in online live classrooms. Such classrooms have been gaining significant attention for K-6 students in recent time[1−3].

• Yuanyi Zhen and Jar-Der Luo are with the Laboratory of Computational Social Science and National Governance, Tsinghua University, Beijing 100084, China. E-mail: zheny21@mails.tsinghua.edu.cn; jarderluo@126.com.
• Hui Chen is with the School of Chinese Language and Literature, Beijing Foreign Studies University, Beijing 100089, China. E-mail: chenhui@bfsu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2023-02-13; revised: 2023-06-01; accepted: 2023-06-06

Online classrooms are similar in format to public schools, where one teacher gives a lecture and students listen, but there are some critical differences, including personalized instruction and open interactions in both teacher-student and student-student patterns in chat rooms[4]. Students construct knowledge through chained sequences of utterances, which contains the logic of thinking and query of students[5].

However, students may achieve different levels of academic performance in the same class. Thus, recognizing the different academic performance levels of students and giving them timely, effective feedback and guidance is a critical issue. Several categories of information can be detected from classroom dialogues, including emotion, cognition, and behavior[6], each have different effects on the academic performance of students. Meanwhile, rich availability of learning data of online classrooms and deep learning tools provide a new means to automatically identify interaction features and build prediction models for academic performance[7, 8]. This study aims to take a big data-based approach to fill the gap in existing studies on the features of classroom dialogues and developing ways for dialogue classification and academic performance prediction.

Our research is carried out with a large dataset of online classrooms from a large educational technology company in China. More specifically, we seek to leverage natural language processing and deep learning models on a large amount of online live classroom dialogue data to reveal the specific relationship between students' attributes in classroom dialogue and their academic performance. We first train text classification models to automatically recognize types of dialogue. Related features in dialogues are extracted and used for identifying the behavioral patterns by which high-performing students are distinguished from their low-performing counterparts. To do that, we employ decision tree, artificial neural network, and convolutional neural network models to identify the effectiveness of academic performance prediction and then build an interpretable model for the best prediction model. The main research questions are as follows:

(1) Is there any difference in interactive types between high-and low-performance groups in both STEM and non-STEM courses?

(2) Is there any difference in students' emotion shown in class dialogues between high- and low-performance groups in both STEM and non-STEM courses?

(3) What are the most important predictors for the best performance prediction model in STEM and non-STEM courses, respectively?

2 Literature Review

2.1 Classroom dialogue and academic performance

As a widely used method for learning and teaching, classroom dialogue was defined as "when one individual addresses another individual or individuals and at least one addressed individual reply"[9]. A productive classroom dialogue should be achieved in a collective, reciprocal, supportive, regulated, and dynamic way[10]. Knowledge will be constructed through these chained sequences of utterances, which contain the logic of thinking and inquiry of students[5]. Since 1970s, many studies have been focused on verifying the importance of classroom dialogue for primary school students[9], including developing critical thinking and reasoning ability[11], and improving learning achievement[9, 11, 12].

Classroom interaction contains different meanings, including emotional, cognitive, and behavioral[6, 13, 14]. The interactive types and emotional expression are critical aspects of classroom dialogue, which have different functions for student learning. Below, we describe the affordances of emotional and interactive aspects of classroom dialogue in terms of learning.

2.1.1 Emotion and learning

Emotion can be viewed as a complex set of components consisting of affective experiences, cognitive processes, physiological adjustment, and behavioral tendencies[15, 16]. The emotional expression of classroom dialogue can be classified into two sub-types: positive and negative, each being further divided into emotional response, evaluation, and expression[17, 18]. The positive emotional expression includes joy, enthusiasm, excitement, and pride; and the negative emotional expression includes fear, anxiety, stress, and guilt[6]. Emotional behavior is one of the critical components related to knowledge construction and creation[13], which can influence cognitive resources[19], learning motivation, and learning strategy[20]. This can either impede or motivate learning[21], so it has a significant impact on the

academic performance of students.

The relationship between emotion and learning in the classroom has received much investigation[22]. As for positive emotion, most research has shown that they can promote learning by way of helping students maintain cognitive resources, focusing on learning tasks, activating intrinsic motivation and learning interest, as well as facilitating deep learning[23]. As for negative emotion, there are two kinds of views. On the one hand, negative emotion undermines learning[24, 25], which can reduce cognitive resources, distract from task-related attention, undermine both intrinsic and extrinsic motivation, as well as promote shallow information processing by students[23]. On the other hand, some studies suggest that negative emotion can benefit learning with the guidance of the teacher[25, 26]. This is especially in terms of the feeling of confusion, which can be a marker of conflict between information stream and knowledge construction, and is related to academic performance[27]. Negative emotion (anger, anxiety, and shame) can also trigger extrinsic learning motivation to avoid failure by striving to do better[23].

2.1.2 Interactive type and learning

Transcripts from classroom dialogue can be classified into three interactive types: cognitive, metacognitive, and off-topic[28]. Cognitive interaction refers to the interaction with others related to knowledge and ideas, including raising doubts, providing constructive feedback and knowledge building, which is beneficial for individual and collective knowledge advancement[14]. Students have off-topic dialogues unrelated to the academic task they have been given when they have some free time or the teacher is absent. It distracts students' ability to focus on tasks and decreases the knowledge convergence for the whole class in the classroom. However, for classes of project-based learning, students will work better when there is more off-topic chit-chat and general socializing. Off-topic interaction may be a catalyst for students to form new and creative ideas[29].

Metacognition of students has been defined as the monitoring and control of cognition[30]. Metacognition can be beneficial for students' problem-solving and critical thinking skills[31]. Metacognitive activities of students in the classroom include planning, monitoring, and evaluation[32], which can monitor and regulate the learning state and motivate students to focus so as to be involved in the classroom[33, 34]. Some studies have illustrated that metacognition is a powerful positive predictor of students' academic performance[35−38]. Few works focus on the effect of dialogue attributes on academic performance during a specific lesson stage, which is one of the research goals of this study.

2.1.3 Lesson stages

Various studies have been conducted on different stages of classroom lessons, which have different teaching purposes and students may display different behavior or attention levels. Most of such studies divided lesson stages by teaching activity type[39−42]. The relationship between different lesson stages and students' emotional behavior received significant attention. For example, Tonguç and Ozkara analyzed emotional change and its relationship with achievement in three different stages: introduction, activities, and closure[43]. In this research, due to the large number of courses and their substantial differences in teaching activities, it is not suitable to divide the learning stages guided by teaching activities. The method of three-stage lesson segmentation was adopted in our study, including the beginning (opening and objective), middle (input and guided practice), and summary stage (review, evaluation, and homework assignment).

2.2 Automatic classification of classroom dialogue text

Many studies have focused on the automatic classification of class structures or activities based on classroom dialogue[44−47]. Yet, most of such studies ignore the importance and meaning of dialogue which results in the quality of the in-class interaction not being measured. Meanwhile, there have been few studies focusing on semantic content to achieve the automatic classification of primary school classroom dialogue. For example, Song et al.[48] automatically classified textual classroom discourse based on the semantic content into seven kinds (prior-known knowledge, analysis, coordination, speculation, uptake, agreement, and querying). The algorithm of Bert and CNN-BiLSTM were used in this research and the overall F1 score is 68%. It is rare for existing studies to automatically detect emotional expression and interactive types based on semantic content simultaneously. In addition, many studies ignored student discourse in student-student dialogue and paid more attentions to teacher-student interaction. This is because that data are so hard to collect in the traditional

classroom.

To overcome these limitations, the current research aims to automatically classify emotional expression and interactive types in the live classroom based on the BERT model, which is a model in natural language processing tasks, which for example, can perform text classification[8].

2.3 Academic performance prediction and interpretability

Some studies have been conducted on how to predict at-risk students in K-6[1, 2, 4], they focused on the features of key courses, attendance, suspension, drug use, behavior, and course performance. While some academic performance predictions are focusing on forum text, features are extracted from the length and sentiment type of the post[49, 50] and intention identification[7]. Classification algorithms include the CNN-LSTM model and Gradient Boost Regression, and prediction results can reach 60%−90%[49, 50]. However, few academic performance prediction studies have focused on K-6 students in online courses, especially in a live classroom. Further, they focused on the identification of at-risk students, while it is also critical to identify the students with different levels of learning acquisition. To our knowledge, analyzing classroom dialogue to better understand academic improvement prediction in the live classroom has received little attention to date.

Machine learning can address a wide range of prediction problems and obtain excellent prediction results in education domains. However, due to the complexity of internal functions in prediction models[51], decision-makers and teachers cannot understand the calculation process of models and therefore do not easily trust the results they provide. Meanwhile, some interpretable AI tools are used to improve general prediction model's interpretability[52]. There are two model-agnostic interpretable approaches for deep learning algorithms[53]. One is LIME (local interpretable model-agnostic explanations), which can provide local explanations for prediction models based on local assumptions[54]. Another one is SHAP (Shapley additive explanations), which can offer global explanations based on game theory and can calculate Shapley values (the contribution to the difference between the prediction power of a specific model and the average predicted value) to measure the contributions of each feature in a feature set[55]. Some studies explored the explanations of grade or knowledge prediction models using SHAP[56, 57]. In order to gain a better understanding of prediction results, we choose SHAP as an interpretable AI tool for our academic prediction to identify the most important predictors.

3 Method

The methodology of this study is depicted in Fig. 1. The study was conducted using five steps: data

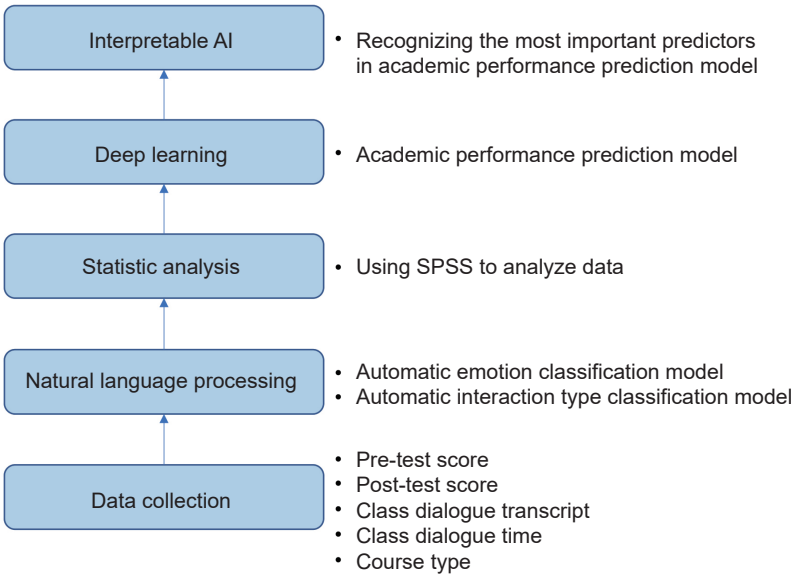


Fig. 1 Methodology for data collection and analysis.

collection from the database, natural language processing for classroom dialogue, statistical analysis for classroom dialogue indicators, using deep learning algorithms to build academic performance prediction model, and using an interpretable AI model to recognize the most important predictors in the trained prediction model. More detailed descriptions for each main step are provided in the following sections.

3.1 Dataset

The data are collected from the live classroom in a large online learning platform in China. In a live classroom, one student sends a message that can be seen by all students in the classroom and they can send timely responses in a chat room. We collected a live classroom dialogue transcript of different subjects and grades from the platform's K-6 courses. According to the result of the spring semester in the year 2020, there were 122 324 K-6 students. A total of 32 630 students participated in 10 179 non-STEM courses while 89 694 students participated in 15 189 STEM courses, which generated 2 619 816 and 6 762 196 interactive texts, respectively. Table 1 shows the distribution of live classroom dialogue text in the chat room. *M* stands for mean value.

The measure of students' academic performance in our study is the difference between the pretest rank and the posttest rank. We extract the top 20% in academic performance rank as the high-performing group and the bottom 20% as the low-performing group. 19 103 students (9837 low-performing students and 9266 high-performing students) for non-STEM courses and 54 634 students (30 842 low-performing students and 23 792 high-performance) for STEM courses are included in the following analyses.

3.2 Natural language processing

In order to automatically recognize the emotional expression and interactive types of classroom dialogue, we have trained two text class ification models. The flow chart can be seen in Fig. 2. The following sections will illustrate the main steps in detail.

3.2.1 Classification schema of classroom dialogue texts

We set the classification standard for emotional expression and interactive types, and give the corresponding text examples from the class. For emotional expression, we can divide dialogue text into positive or negative emotion. If students express joy, happiness, and excitement, it is categorized as positive emotion. While if they express sadness, boredom, and anger, it is defined as negative emotion. Examples can be seen in Table 2.

For interactive types, we choose three kinds of classification standards: cognition, meta-cognition, and off-topic[28]. Cognition includes knowledge building, asking questions to teachers or classmates and responding. Meta-cognition includes planning, monitoring, reflection, and evaluation, which reflects the regulation of learning processes in the classroom.

Table 1 Distribution of live classroom dialogue text.

Course type	Subject	Grade number	Number of classes	*M* (SD)		Number of interactive texts in course
				Number of students in class	Number of interactive texts in student	
Non-STEM course	English	1	637	1.64 (1.11)	87 (107)	90 670
		2	929	5.15 (3.8)	57 (56)	272 484
		3	2443	2.6 (1.9)	82 (76)	527 474
		4	1532	3.5 (2.5)	88 (86)	474 405
		5	2048	3.3 (2.8)	84 (85)	570 212
		6	2029	3.4 (2.5)	80 (97)	559 771
	Chinese	2	1	1	22	22
		6	560	2.16 (1.33)	103 (147)	124 778
STEM course	Math	1	2647	7.7 (6.1)	59 (46)	1 209 799
		2	4235	6.5 (5.7)	55 (51)	1 525 599
		3	2959	5.4 (4.4)	77 (69)	1 235 707
		4	1666	3.7 (3.2)	96 (84)	584 828
		5	2160	5.2 (3.3)	110 (108)	1 222 418
		6	1522	5.4 (3.6)	119 (134)	983 845

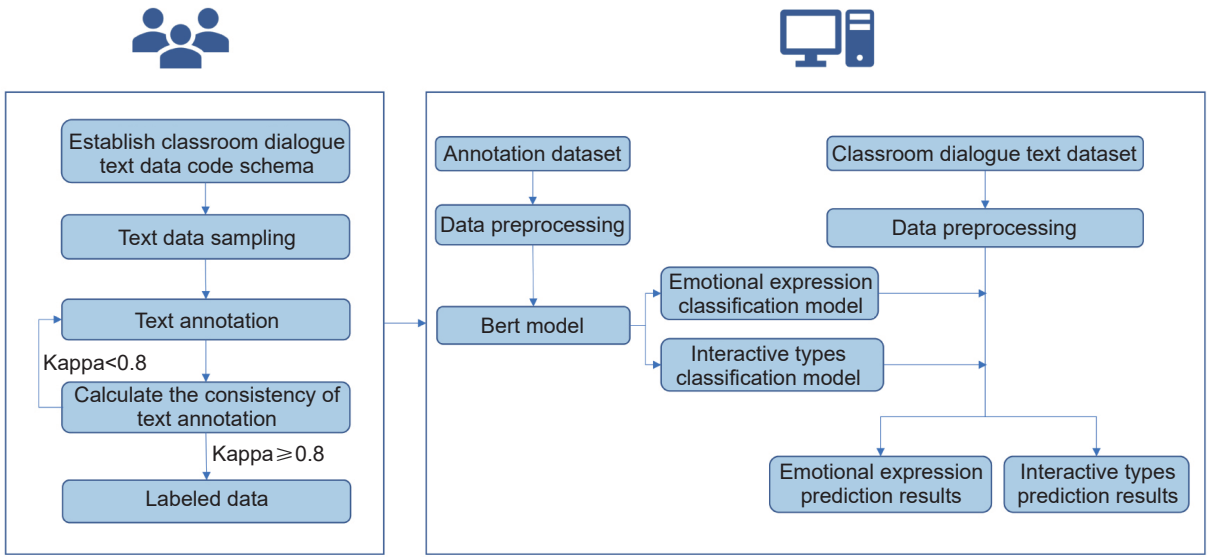


Fig. 2 Flow chart of the prediction of classroom dialogue classification.

Table 2 Coding scheme of live classroom dialogue text.

Dimension	First-level category	Second-level category	Example
Emotional expression	Positive	—	"It's so exciting to hear this story in English."
	Negative	—	"It's so boring."
Interactive type	Cognition	Knowledge building	"The imaging principle of concave lens is different from that of convex lens."
		Ask questions	"What is a common factor?"
		Answering	"The answer is 'books'"
	Meta-cognition	Planning	"The first part of this class will be taught by the teacher and the second part we will study by ourselves."
		Monitoring	"The teacher will come back soon."
		Reflection and evaluation	"I still don't understand this part and I should listen more carefully."
	Off-topic	Off-task statements and social greetings	"I'm so hungry." "Hello everyone, I'm Li Jun" Emoji

Off-topic dialogue occurs when students express something unrelated to learning or engage in social greetings. Specific examples can be seen in Table 2.

3.2.2 Manual annotation

According to a certain percentage of dialogue texts for every course in the whole dataset, as shown in Table 1, we sample 4 540 230 and 10 946 pieces of dialogue texts from English, Chinese, and Math courses, respectively.

Based on the above classification scheme of classroom dialogue texts, two data annotators are invited to manually classify 5% of the randomly selected text data. Typically, two independent annotations are necessary if they are in agreement on a same result, otherwise a third one is introduced. And the Kappa coefficient was used to verify the consistency and reliability of the classification results. After multiple rounds of iterative modification, the kappa values of emotional expression and interactive types are 1.000 (p<0.001) and 0.964 (p<0.001), respectively, which mean two data annotators are highly consistent in text classification standards. Therefore, text annotation is completed by these data annotators, and each annotated 7808 texts.

3.2.3 Classroom dialogue text classification model

The model of Bidirectional Encoder Representations from Transformers (BERT) is employed to train and predict the emotional expression and interactive types based on annotated data. BERT model is pre-trained in unlabeled text and fine-tuned in experimental data with one output layer, which has been applied in natural language processing tasks, such as text classification[8]. The model is constituted by a multi-layer bidirectional transformer encoder, which can reduce model training

time by paralleling the calculation and adding the attention mechanism. There are two parts of the BERT model, including Pre-training and Fine-Tuning. In the Pre-training part, the word vector is trained by way of a masked model and next sentence prediction. And experimental data are input and trained in the Fine-Tuning part[58].

We train and evaluat our model in a single machine with GeForce GTX 1080 Ti GPU, and choose the Pre-training model as Bert_chinese_L-12_H-768_A-12. Then the model sets the max sequence length as 256, training batch size as 16 and learning rate as 10^{-5} in the Fine-Tuning part.

3.2.4 Evaluation of dialogue classification model of emotional expression and interactive types

The indicators of Precision (*P*), Recall (*R*), and F1 are used to evaluate the classification results in emotional expression and interactive types. For the emotional expression classification model, the prediction result is shown in Table 3. The accuracy (ACC) reached 96.4% and the F1 score of predicting positive emotion text and negative emotion text is 98.1% and 62.6%, respectively.

For the interactive type classification model, the accuracy is 91.4%, the F1 score of cognition text, meta-cognition text, and off-topic text are 93.9%, 48%, and 92.8%, respectively, which can be seen in Table 3. The prediction ability of the model for cognition and off-topic texts is better than the meta-cognition information.

3.3 Relationship between dialogue text and academic performance

Because the labels of classroom dialogue classification cannot be calculated directly, they should be transformed into explanatory variables for students' performance in the next step. As the class progresses, the students' emotional expression and interactive types keep changing. For example, at the beginning of class, students are usually positive and focus on knowledge cognition; in the middle phase of class, students usually focus on cognitive and meta-cognitive aspects; towards the end of class, students are more tired, tend to feel negative about being in class, and engage in off-topic dialogue. In order to obtain more informative data reflecting in-class behaviors, the whole class is divided into three time slots, beginning with the time the first sentence was sent by a student and ending with the time of the last sentence sent by a student in the chat room. We split the three phases evenly during the period between start time and end time. We labeled the stage of each dialogue text as the beginning stage, middle stage, and summary stage in each class according to the time the text was sent, respectively.

In order to recognize the interaction style of a student, we compute the proportion of the number of texts for each type of emotional expression, interactive types at the individual level, and the individual share of the overall text numbers in each stage of the whole class.

The variables extracted from in-class live classroom dialogue by the above-stated data processing can be categorized into three dimensions: student interaction styles, interaction features (emotional expression, interactive types, and interaction frequency), and class stages. Each variable is named after three letters to define the features of the three dimensions as Table 4.

All the variables of features extracted from in-class dialogue text are further explained in Table A1 of Appendix.

Table 3 Evaluation of Bert model for live classroom dialogue text classification.

Indicator		*P*	*R*	F1	ACC
Emotional expression	Positive	0.968	0.994	1	0.964
	Negative	0.836	0.5	0.6	
Interactive type	Cognition	0.936	0.941	0.9	0.914
	Meta-cognition	0.513	0.451	0.5	
	Off-topic	0.926	0.931	0.9	

Table 4 Definitions of feature dimensions and variables.

Dimension	Letter	Description
Student interaction style	I	Proportion of a certain emotion or type of a student's interaction among all his/her interactions.
	C	Certain emotion or type of in-class interactions, the individual share of whole class's interactions.
	P	Positive emotional expression in interactions.
Interaction feature (emotion, types, and frequency)	N	Negative emotional expression in interactions.
	C	Cognition type of interaction.
	M	Meta-cognition type of interaction.
	O	Off-topic type of interaction.
	T	Frequency of interaction measured by the number of sending message.
Lesson stage	B	Beginning stage of a class.
	M	Middle stage of a class.
	S	Summary stage of a class.
	A	All stages of a class.

To reveal the difference between in-class interaction features and academic performance level, Mann-Whitney *U* test is performed to test the significance of difference between the two groups as the data distribution, which is not consistent with the normal distribution.

3.4 Academic performance prediction

Features can be divided into two parts: the first one is performance before class (pretest rank) which indicates the basic learning level of students, and the other one is features extracted from classroom dialogue text, which can be seen in Table A1 of Appendix. In total, there are 48 features in the feature set to construct prediction models.

We compare three classification algorithms for academic performance prediction, including decision tree (DT), artificial neural network (ANN), and convolutional neural network (CNN). We compare three evaluation indicators, including recall, precision, and accuracy for the three methods to select the best prediction model, and then construct an interpretable model. TensorFlow deep learning framework is used in our study to build the prediction model.

3.5 Interpretable AI

The library of Shap in python is used to construct interpretable models. Different methods can be used for the SHAP library to calculate Shapely value depending on the prediction algorithms. We use the tree-based approach (TreeSHAP) to calculate feature contributions for the decision tree algorithm[55]. DeepSHAP is used to explain feature contributions of deep learning algorithms, such as deep neural networks and convolutional neural networks. The SHAP visualization pictures can show the predictors' impact on prediction results, and the attribute is ranked by its contribution from top to bottom. We can also understand the impact direction by color (red is associated with a positive impact, while blue means negative one).

4 Result

4.1 Difference between performance groups for the emotional expression

Non-parametric Mann-Whitney *U* test was used to explore whether two groups of students (high level versus low level) in STEM courses and non-STEM courses, respectively, differed in regards to their emotional expression. Tables 5 and 6 show the comparison of four features between the high- and low-performing groups in the three stages and during the whole class as follows: (1) proportion of positive emotion at the individual level, (2) proportion of negative emotion at the individual level, (3) individual share of collective positive emotion, and (4) individual share of collective negative emotion.

4.1.1 Non-STEM courses

In non-STEM courses, there are significant differences in the individual share of collective positive emotion in the beginning, middle, and summary stages, and during the whole class between high- and low-performing students. That is, high-performing students tend to express high share of positive emotional expression of the whole class in all stages of the lesson. But in the summary stage, high-performing students prefer to express higher share of negative emotion in the whole class than the low-performing group.

4.1.2 STEM courses

Table 6 depicts the difference in emotional expression features between high- and low-performing students in

Table 5 Difference between high- and low-performance groups in emotional expression features of non-STEM courses.

Feature	*M* (SD)		*U*
	Low-performing students in non-STEM courses (*N*=9837)	High-performing students in non-STEM courses (*N*=9266)	
IPA	0.984 (0.037)	0.984 (0.038)	4.54×10^{7}
CPA	0.312 (0.312)	0.335 (0.314)	4.301×10^{7}***
INA	0.016 (0.037)	0.016 (0.038)	4.54×10^{7}
CNA	0.214 (0.349)	0.217 (0.348)	4.51×10^{7}
IPB	0.960 (0.157)	0.963 (0.151)	4.54×10^{7}
CPB	0.311 (0.315)	0.335 (0.319)	4.322×10^{7}***
INB	0.015 (0.041)	0.015 (0.041)	4.52×10^{7}
CNB	0.166 (0.332)	0.171 (0.333)	4.51×10^{7}
IPM	0.910 (0.265)	0.924 (0.243)	4.52×10^{7}
CPM	0.297 (0.316)	0.322 (0.319)	4.291×10^{7}***
INM	0.015 (0.057)	0.014 (0.052)	4.52×10^{7}
CNM	0.115 (0.289)	0.120 (0.294)	4.51×10^{7}
IPS	0.939 (0.210)	0.952 (0.183)	4.54×10^{7}
CPS	0.309 (0.319)	0.332 (0.320)	4.287×10^{7}***
INS	0.016 (0.050)	0.015 (0.052)	4.51×10^{7}
CNS	0.143 (0.318)	0.151 (0.324)	4.494×10^{7}*

Note: ***$p < 0.001$ and *$p < 0.05$.

Table 6 Difference between high- and low-performance groups in emotional expression features.

Feature	*M* (SD)		*U*
	Low-performing students in STEM courses (*N*=30 842)	High-performing students in STEM courses (*N*=23 792)	
IPA	0.986 (0.030)	0.987 (0.029)	3.610×10^{8}***
CPA	0.183 (0.214)	0.200 (0.229)	3.507×10^{8}***
INA	0.014 (0.030)	0.013 (0.029)	3.610×10^{8}***
CNA	0.144 (0.268)	0.145 (0.275)	3.638×10^{8}*
IPB	0.957 (0.171)	0.954 (0.181)	3.66×10^{8}
CPB	0.183 (0.218)	0.200 (0.234)	3.519×10^{8}***
INB	0.014 (0.040)	0.013 (0.038)	3.66×10^{8}
CNB	0.119 (0.267)	0.123 (0.272)	3.67×10^{8}
IPM	0.935 (0.225)	0.934 (0.227)	$3.66\times10^{8}08$
CPM	0.180 (0.219)	0.197 (0.234)	3.520×10^{8}***
INM	0.013 (0.045)	0.012 (0.041)	3.66×10^{8}
CNM	0.089 (0.243)	0.089 (0.245)	3.66×10^{8}
IPS	0.954 (0.183)	0.958 (0.174)	3.624×10^{8}***
CPS	0.182 (0.219)	0.201 (0.235)	3.507×10^{8}***
INS	0.012 (0.043)	0.011 (0.041)	3.638×10^{8}**
CNS	0.100 (0.258)	0.097 (0.257)	3.643×10^{8}*

Note: ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$.

STEM courses. Similar to non-STEM courses, in the individual share of collective positive emotion, in the beginning, middle, and summary stages as well as during the whole class, STEM courses' high-performing students express more positive dialogue than those low-performing ones. In addition, for positive emotion, high-performing students have a significantly higher rank improvement regarding the proportion of positive emotion at the individual level in both the summary stage and during the whole class. In contrast with non-STEM courses, the high-performing group shows lower negative emotion than its counterpart at both the individual and collective level of the summary stage.

In summary, for both non-STEM and STEM courses, high-performing students always express more positive dialogue in all stages of the lesson than low-performing students. High-performing students in non-STEM courses show more negative dialogue in the summary stage, which can be seen in low-performing students in STEM courses.

4.2 Difference between performance groups for the interactive type

4.2.1 Non-STEM courses

In non-STEM courses, most features show significant differences between high-performing and low-performing students, except for the proportion of off-topic at the individual level in middle and summary stages, which can be seen in Table 7. For high-performing students, there are significant negative impacts regarding the proportion of off-topic dialogue in their all-class interactions in all three lesson stages and during the whole class. For the other two interactive types at the individual level and all three interactive types at the collective level, high-performing students show a positive impact, i.e., higher improvement in in-class ranking than the low-performing group in the beginning, middle, and summary stages as well as during the whole class.

4.2.2 STEM courses

The results of the difference test for STEM courses are presented in Table 8. For cognitive dialogue, except for

Table 7 Difference between high- and low-performance groups in interactive types of non-STEM courses.

Feature	*M* (SD)		*U*
	Low-performing students in non-STEM courses (*N*=9837)	High-performing students in non-STEM courses (*N*=9266)	
ICA	0.472 (0.199)	0.482 (0.195)	4.434×10^{7}***
CCA	0.307 (0.311)	0.335 (0.314)	4.243×10^{7}***
IMA	0.027 (0.044)	0.028 (0.044)	4.402×10^{7}***
CMA	0.245 (0.347)	0.265 (0.352)	4.367×10^{7}***
IOA	0.500 (0.192)	0.490 (0.187)	4.413×10^{7}***
COA	0.313 (0.316)	0.330 (0.319)	4.366×10^{7}***
ICB	0.457 (0.240)	0.469 (0.236)	4.422×10^{7}***
CCB	0.304 (0.315)	0.332 (0.319)	4.268×10^{7}***
IMB	0.025 (0.051)	0.027 (0.057)	4.440×10^{7}***
CMB	0.205 (0.342)	0.226 (0.354)	4.401×10^{7}***
IOB	0.494 (0.236)	0.482 (0.229)	4.422×10^{7}***
COB	0.311 (0.321)	0.328 (0.326)	4.399×10^{7}***
ICM	0.454 (0.301)	0.473 (0.289)	4.397×10^{7}***
CCM	0.282 (0.317)	0.316 (0.324)	4.210×10^{7}***
IMM	0.021 (0.061)	0.023 (0.064)	4.407×10^{7}***
CMM	0.143 (0.310)	0.162 (0.322)	4.402×10^{7}***
IOM	0.450 (0.296)	0.442 (0.280)	4.50×10^{7}
COM	0.290 (0.320)	0.309 (0.325)	4.365×10^{7}***
ICS	0.443 (0.263)	0.454 (0.250)	4.441×10^{7}**
CCS	0.300 (0.319)	0.329 (0.322)	4.242×10^{7}***
IMS	0.030 (0.070)	0.031 (0.064)	4.414×10^{7}***
CMS	0.191 (0.340)	0.210 (0.350)	4.396×10^{7}***
IOS	0.481 (0.261)	0.482 (0.247)	4.55×10^{7}
COS	0.307 (0.327)	0.327 (0.328)	4.330×10^{7}***

Note: ***$p < 0.001$ and **$p < 0.01$.

Table 8 Difference between high- and low-performance groups in interactive types of STEM courses.

Feature	*M* (SD)		*U*
	Low-performing students in STEM courses (*N*=30 842)	High-performing students in STEM courses (*N*=23 792)	
ICA	0.455 (0.182)	0.462 (0.181)	3.588×10^{8}***
CCA	0.181 (0.213)	0.201 (0.229)	3.479×10^{8}***
IMA	0.028 (0.047)	0.028 (0.050)	3.65×10^{8}
CMA	0.162 (0.261)	0.173 (0.273)	3.64×10^{8}
IOA	0.516 (0.178)	0.510 (0.179)	3.587×10^{8}***
COA	0.183 (0.219)	0.199 (0.234)	3.539×10^{8}***
ICB	0.456 (0.227)	0.459 (0.226)	3.632×10^{8}*
CCB	0.181 (0.217)	0.200 (0.233)	3.496×10^{8}***
IMB	0.028 (0.058)	0.027 (0.058)	3.67×10^{8}
CMB	0.144 (0.268)	0.154 (0.278)	3.634×10^{8}*
IOB	0.487 (0.224)	0.480 (0.223)	3.606×10^{8}***
COB	0.183 (0.226)	0.199 (0.240)	3.544×10^{8}***
ICM	0.425 (0.273)	0.428 (0.270)	3.64×10^{8}
CCM	0.176 (0.223)	0.195 (0.238)	3.507×10^{8}***
IMM	0.025 (0.070)	0.025 (0.071)	3.66×10^{8}
CMM	0.119 (0.263)	0.123 (0.270)	3.65×10^{8}
IOM	0.497 (0.278)	0.493 (0.275)	3.628×10^{8}*
COM	0.179 (0.225)	0.193 (0.239)	3.555×10^{8}***
ICS	0.438 (0.251)	0.446 (0.246)	3.594×10^{8}***
CCS	0.179 (0.220)	0.200 (0.237)	3.477×10^{8}***
IMS	0.028 (0.066)	0.028 (0.068)	3.65×10^{8}
CMS	0.135 (0.272)	0.142 (0.281)	3.64×10^{8}
IOS	0.500 (0.255)	0.495 (0.248)	3.616×10^{8}**
COS	0.182 (0.226)	0.199 (0.242)	3.538×10^{8}***

Note: ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$.

the proportion of cognition at the individual level in the middle stage, high-performing students show significantly greater improvement in ranking than low-performing students due to the high level of interaction at both individual and collective levels. For meta-cognition dialogue, only in the beginning stage high-performing students are significantly impacted positively by the individual share of collective meta-cognition, while all other meta-cognition dialogue of both individual and collective types in all the stages had either a negative or insignificant influence. For off-topic information, high-performing students are significantly impacted negatively by the proportion of off-topic chat in the individual in-class dialogue, but influenced positively by the individual share of the whole class's interaction.

In summary, for both non-STEM and STEM courses, we found that high-performing students prefer engaging in more cognition and off-topic dialogue in all stages of the lesson. But a high proportion of off-topic dialogue in individual-level interaction generates a negative effect on academic performance. High-performing students in non-STEM courses express more meta-cognition information, while this cannot be seen in STEM courses.

4.3 Predicition of the academic performance of non-STEM courses

Following the significant differences between high- and low-performing students' in-class interaction observed by the Mann-Whitney *U* test, DT, ANN, and CNN are used to predict two groups' academic performance for both non-STEM courses and STEM courses.

In non-STEM courses, the C1 group contains only features related to online live classroom interactions and the C2 group is the combination of C1 group and the pretest rank feature. For the C1 group, the optimal results were achieved using a DT with max_features=4, min_samples_split=2, and min_samples_leaf=1. Additionally,it had a deep ANN with two hidden layers of 240 and 32 neurons, a batch size of 128, and a learning rate of 0.000 497, as well as a CNN with a reshape layer and two convolutional layers of 32 and 16 filters with a kernel size of (3, 3) and ReLU activation function.

The C2 group, with DT configured using max_features=26, min_samples_split=163, and min_samples_leaf=84, an ANN with three hidden layers of 208 480, and 400 neurons, a batch size of 128, and a learning_rate of 0.00 377, and a CNN employing a reshape layer and three convolutional layers with 16 filters each, using a kernel_size of (3, 3) and the ReLU activation function, resulted in optimal predictive outcomes for each model.

The prediction results can be observed in Table 9, and the CNN model in the C2 group shows the best prediction ability with an accuracy of 84%. The effective features of the performance prediction model are analyzed by SHAP, which is a method of interpretable artificial intelligence. The results can be seen in Fig. 3. Meanwhile, the C1 group shows little improvement in accuracy compared to the baseline (50%). This implies that online classroom interaction features have little significant impact on non-STEM courses.

Table 9 Evaluation results of the academic performance prediction model.

Group	Category	Model	Type	*P*	*R*	F1	ACC
C1	Non-STEM course	DT	Low	0.53	0.52	0.53	0.51
			High	0.49	0.50	0.50	
		ANN	Low	0.54	0.59	0.56	0.52
			High	0.50	0.45	0.47	
		CNN	Low	0.54	0.49	0.52	0.52
			High	0.50	0.56	0.53	
	STEM course	DT	Low	0.58	0.67	0.62	0.53
			High	0.44	0.35	0.39	
		ANN	Low	0.58	0.87	0.69	0.56
			High	0.48	0.16	0.24	
		CNN	Low	0.58	0.85	0.69	0.56
			High	0.47	0.17	0.25	
C2	Non-STEM course	DT	Low	0.83	0.82	0.82	0.82
			High	0.81	0.82	0.82	
		ANN	Low	0.86	0.79	0.82	0.83
			High	0.80	0.86	0.83	
		CNN	Low	0.82	0.87	0.85	0.84
			High	0.86	0.81	0.83	
	STEM course	DT	Low	0.86	0.81	0.84	0.82
			High	0.77	0.83	0.80	
		ANN	Low	0.83	0.86	0.85	0.82
			High	0.81	0.77	0.79	
		CNN	Low	0.84	0.85	0.85	0.83
			High	0.80	0.79	0.80	

Pretest rank in course, the proportion of cognition dialogue at the individual level during the whole class, and individual share of collective off-topic during the whole class are the three most important predictors of the prediction model. In addition, the interactive type of the summary and middle stages are the most and almost equally important two predictors, and it is clear that interactive type is more important than emotional expression.

4.4 Prediction of the academic performance of STEM courses

The prediction of students' academic performance is converted to a binary classification problem by defining two sets of categories, high-performing and low-performings students.

In STEM courses, for the C1 group, optimal results are obtained by employing the following models: DT with max_features=17, min_samples_split=2, and min_samples_leaf=54; an ANN consisting of four hidden layers with 496, 32, 32, and 32 neurons, respectively, a batch size of 128, and a learning rate of 0.000 487 8; and a CNN with a reshape layer, three convolutional layers of 32, 16, and 16 filters, a kernel size of (3, 3), and the ReLU activation function.

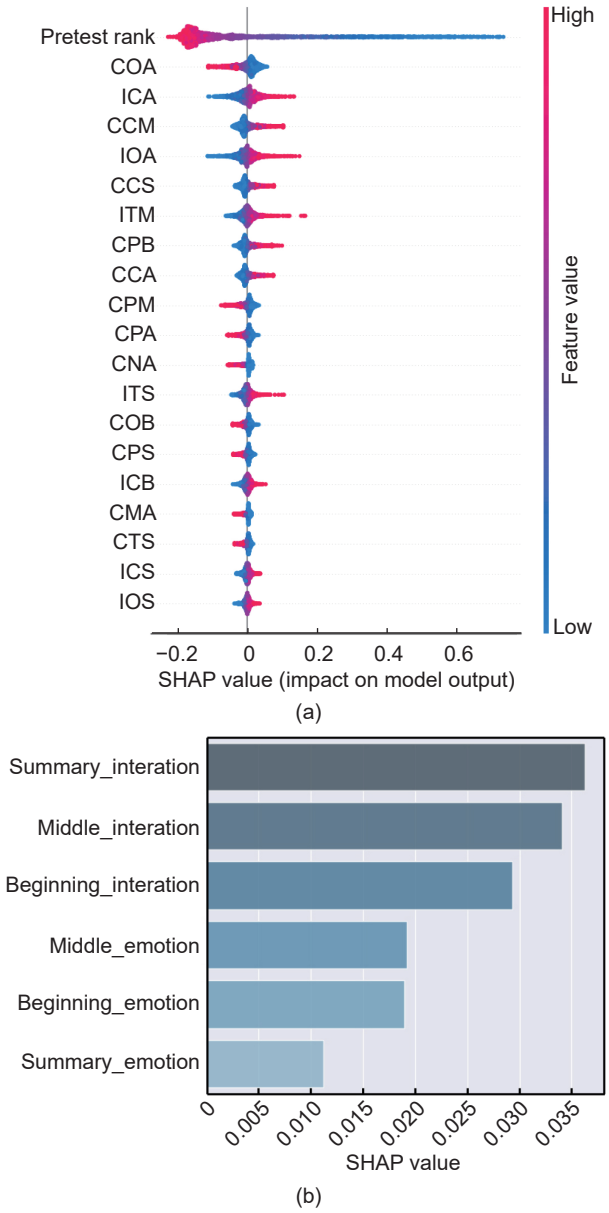


Fig. 3 Interpretable results of the academic performance prediction model in non-STEM courses. (a) The importance rank of each feature in prediction model. (b) The importance rank of interactive types and emotional expression in beginning, middle and summary stages in prediction model. There are six variables, each obtained by summing the absolute values of the interactive types/emotional expression variables in the responding stage, for example, summary_interaction (the SHAP value of interactive types in summary stage) = |ICS|+|IMS|+|IOS|+|CCS|+|CMS|+|COS|, summary_emotion (the SHAP value of emotional expression in summary stage) = |IPS|+|CPS|+|INS|+|CNS|.

Regarding to the C2 group, DT with max_depth is 4, max_features is 34. Deep ANN with four hidden layers of 208, 32, 32, and 32 neurons, dropout is 0.25, learning_rate is 0.000 134 and a batch size of 128. CNN with reshape layer, three convolutional layers of 32, 16, and 16 filters, kernel_size is (3, 3) and the activation function is "Relu", produced optimal results.

The prediction results can be observed in Table 9, CNN model in the C2 group performs the best prediction ability that accuracy is 83% and the interpretable results of the prediction model can be seen in Fig. 4. Furthermore, there has been a 6% improvement in accuracy compared to the baseline (50%), suggesting that online classroom interaction holds relatively significant importance for STEM courses compared to non-STEM courses.

Attributes pretest rank in course, the proportion of off-topic at individual level in middle stage, and the proportion of cognition dialogue at individual level in middle stage are the most effective predictors for recognizing two-level students in STEM courses. Besides, the interactive type in the middle stage is the most important predictor of academic performance. Similar to non-STEM courses, interactive type is more important than emotional expression.

The prediction accuracy of variables related to online live classroom interaction has shown a modest but noticeable improvement of 2% − 6% percent compared to the baseline. Although the improvement may not be considered statistically significant, it holds significant implications within the context of China's education system, where student performance is heavily determined by scores. Even a slight enhancement in prediction accuracy for a single course, when applied to student guidance, can have a meaningful impact. If extended to all subjects, it has the potential to enhance students' learning outcomes and significantly influence subsequent academic examinations and future educational opportunities.

Meanwhile, the outcomes of this prediction also shed light on the pivotal role played by students' pretest ranking in determining their academic progress or decline. These scores encompass not only their prior achievement but also reflect their learning habits, styles, attitudes, and other pertinent factors. Consequently, our future research endeavors should delve more deeply into the nuanced analysis of interaction quality, moving beyond mere quantitative aspects such as the number of interactions or the responses received from instructors and students in similar textual contexts among different students. Incorporating these significant factors will facilitate a more comprehensive understanding of the impact of live online classroom interactions on students'

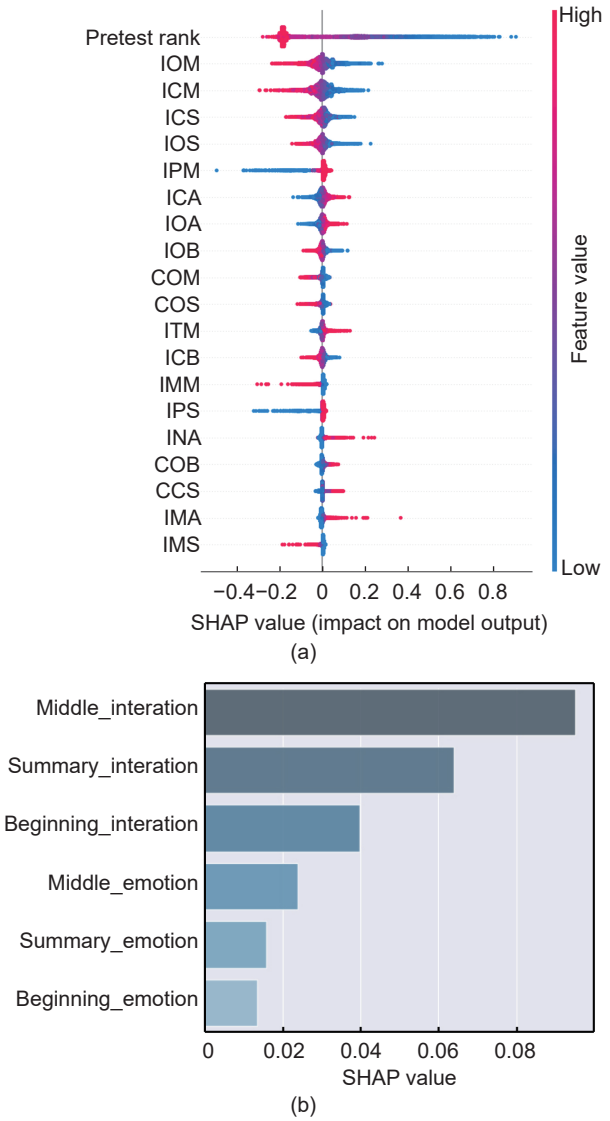


Fig. 4 Interpretable results of the academic performance prediction model in STEM courses. (a) The importance rank of each feature in prediction model. (b) The importance rank of interactive types and emotional expression in beginning, middle and summary stages in prediction model. There are six variables, each obtained by summing the absolute values of the interactive types/emotional expression variables in the responding stage, for example, summary_interaction (the SHAP value of interactive types in summary stage) = |ICS|+|IMS|+|IOS|+|CCS|+|CMS|+|COS|, summary_emotion (the SHAP value of emotional expression in summary stage) = |IPS|+|CPS|+|INS|+|CNS|.

academic performance. Moreover, this heightened level of analysis necessitates increase rigor in data collection, encompassing a broader spectrum of variables and contexts.

5 Discussion and Conclusion

This research aims to discover the relation between classroom dialogue and the academic performance of students in primary school. Two automated text classification models based on the natural language processing techniques are employed to identify interaction features. Then we explore whether interaction features can be used to predict the academic performance of students. Further, we select the best prediction models and build the interpretable artificial intelligence models to identify the most important predictors in them. It is found that some interaction features and pretest rank have a valid predictable power for the academic performance of students.

To combat the challenge of analyzing a large scale of text data, this study automatically recognizes the emotional expression and interactive type of classroom dialogue by two text classification models based on BERT model, a model in natural language processing. Our model achieved better performance in a more challenging text classification task (classifying large-scale text based on semantics)[8]. Features are extracted from emotional expression, interactive types and interaction frequency at the individual level, as well as the individual share of collective number in the beginning, middle, and summary stages of the class from recognized labels of classroom dialogue.

A significance test of difference is conducted to investigate the difference of emotional expression features between high and low-academic performance groups. We find that high-performing students express more positive dialogue in all stages of lesson than low-performing students in the two kinds of courses. This result is in line with Ref. [23].

There is a difference between the two types of course. High-performing students in non-STEM courses show more negative dialogue in the summary stage, while this occurs in low-performing students of STEM courses. The reason for this may be that STEM courses require strong logical thinking and students need to organize material, recognize rules, and comprehend complex structures of information to solve problems[59], which can provide a sense of achievement for students and show positive emotion at the last stage. Meanwhile, low-performing students find it harder to comprehend information and solve such problems and thus they will show negative emotion. As for non-STEM courses, students' academic performance has a strong positive and reciprocal relationship with the time they spend learning[60]. The understanding and memorization cannot be completed during class, and thus high-performing students show negative emotion at the last stage due to the delay in achievability.

The influence of interactive types between high and low academic performance groups is checked in non-STEM courses and STEM courses, respectively. The results show that high-performing students always express more cognition and off-topic dialogue in all stages of the class than low-performing students in these two kinds of courses. This is consistent with the researches[14, 29]. It is noted that off-topic dialogue has a positive relationship with academic performance in the online live classroom, which is in line with project-based learning classes[29]. So that learning environment with rich social interaction opportunities can help students benefit from off-topic dialogue.

The difference between the two types of courses shows that high-performing students in non-STEM courses transfer more meta-cognition information, while this is not observed in STEM courses. The reason for this may be that non-STEM courses require more self-regulation, including planning, monitoring, and evaluating[32], while STEM courses consider more knowledge construction. Thus, teachers in non-STEM courses should pay more attention to fostering the meta-cognitive skills of students.

The prediction model and interpretable AI show that interactive type is more important than emotional expression for both non-STEM and STEM courses. However, interactive types in the summary and middle stages have almost equal effects on students' performance for non-STEM students, while the middle stage is more important than the summary stage for STEM students. The possible interpretation is that memorization is the core of non-STEM courses in primary school, and an effective knowledge summary by teachers in the last stage of the lesson can help

students to memorize knowledge. Thus, students benefit most from active interaction in the summary stage, even though they cannot understand the knowledge taught in the beginning stages. For STEM courses, students need to focus on the logical chain of knowledge, and they will not follow up once get lost in the middle stage, in which asking and answering questions are important to fully understand the whole course for STEM students.

Academic performance prediction models are established, and traditional machine learning and deep learning algorithms are employed. Classroom interaction features are more important for STEM courses than non-STEM courses. The results show that a classification accuracy of around 31 percentage points above the baseline (50% for two-classification), which is at a better level than the study of Ref. [61], which based on internet behavior. Further, there is a similar prediction accuracy around 83%−84% between non-STEM courses and STEM courses.

By the method of interpretable AI, we use SHAP to obtain the most effective predictors of academic performance in non-STEM and STEM courses, respectively. Pretest rank[62] in the course is the most important predictor in both two kinds of courses, which is consistent with current research. Yet we can still see that some different predictors are impacting high-performing students in the two kinds of courses. The proportion of cognition at the individual level during all stages of class and individual share of collective off-topic dialogues during the whole class are critical predictors for non-STEM courses. The proportion of off-topic and cognition dialogues at the individual level in the middle stage are effective predictors for STEM courses. Further, we could see that the interactive types are more important than emotional expression in classroom dialogue, and the middle stage and summary stage are more important than the beginning stage in both types of courses.

In all, interactive dialogue can integrate the process of knowledge building and social interaction of students. While previous studies mostly focused on the formal features of dialogue, this study fills the gap of existing research on the semantic features of large-scale classroom dialogue and academic performance prediction. We find that there are differences and similarities in important factors of academic performance prediction models in non-STEM courses and STEM courses and thus teachers and policymakers should make well-informed decisions based on the different course types.

The present study's findings offer valuable practical insights for teachers, online education platforms, and policymakers. For instructors, there are several key considerations to keep in mind. Firstly, it is essential to tailor teaching strategies to the unique needs of each subject area across the three stages of the classroom. Non-STEM educators should prioritize providing feedback on learning achievability during the summary stage, whereas STEM instructors should aim to repeat the teaching logic chain in the middle stage to ensure students keep pace with the course's rigors. Secondly, cultivating students' self-regulation ability is of utmost importance, particularly for non-STEM teachers, as it allows learners to take ownership of their academic progress and become autonomous learners. Finally, instructors should be mindful of providing space for off-topic discussions in the online live classroom environment, as this can help foster emotional connections and build familiarity among students, which may ultimately enhance their academic performance.

For online education platforms, there are several suggestions. Firstly, there should be an improvement in the strategic detection of off-topic dialogue, as it has a positive relationship with academic performance in the online live classroom. Taking a one-size-fits-all approach to dealing with irrelevant conversations should be avoided. Secondly, different support strategies should be employed for different stages of different courses. For non-STEM courses, the summary and middle stages have equal importance on grades, while for STEM courses, the middle stage is more crucial. Therefore, the platform should strengthen the incentive or supervision measures for the summary stage of non-STEM courses to assist students in their learning. Additionally, the platform design for non-STEM courses should enhance support for metacognitive skills. Last but not least, policymakers should realize that different types of courses correspond to different learning needs and habits of students, which determine

the differences in the learning process. Therefore, evaluations of course and teaching quality need to be more diverse.

There are four limitations in this study, including the fact that the division of the three stages may not be accurate as it is divided equally according to the time from the first interactive text to the last interactive text in the chat room, while the actual teaching time may not reflect this. Future research could involve soliciting input from teachers, who can provide additional teaching knowledge and annotation data to facilitate a more precise analysis of the dynamic development of students' emotions and interactive behaviors. Secondly, there is a need for more fine-grained analyses of classroom dialogue in future research. For example, Zheng et al.[63] classified behavioral engagement of classroom interaction into knowledge-building, regulation, support, and agreement, and also conducted a detailed classification of cognitive engagement into remembering, understanding, applying, and evaluating in an Adobe Photoshop class. However, given the diversity of teaching content, pedagogical approaches, and strategies in massive online learning platforms, the patterns and content of classroom interaction texts may exhibit significant variations. Therefore, developing robust methods for fine-grained recognition and analysis of these texts is a significant research challenge, and represents an important area for future research in this field. Thirdly, our research data were collected from an online education platform in China. The generalizability of our findings in this study can be further tested in other learning contexts in the future, such as traditional in-person classrooms, computer-supported collaborative learning environments, and blended learning classrooms that combine online and offline learning. However, it is important to collect context-specific data such as image and audio data. By leveraging computer vision techniques like facial recognition and gesture analysis, as well as advanced speech processing methods, we can analyze factors like emotional states, interaction patterns, and engagement quality. These insights provide a valuable understanding of students' experiences and behaviors across diverse educational settings. Finally, our research focuses on exploring the correlations between online learning interaction text and learning performance. While we satisfy the criteria of temporal precedence and theoretical plausibility for establishing causal relationships, however we do not address the potential influence of confounding variables, covariates, and mediating factors, which may lead to spurious correlations. This is a study on data mining, aiming to identify significant theoretical value and formulate research propositions. Subsequently, hypotheses can be developed based on these findings, and causal analysis can be conducted using regression methods. This limitation highlights the need for future investigations.

Appendix

Table A1 provides an explanation and description of all the variables extracted from in-class dialogue text.

Table A1 Features set from the classroom dialogue dataset.

Feature	Description
IPA	Individual positive emotion ratio
CPA	Positive emotion ratio in class
IPB	Individual positive emotion ratio in beginning stage
CPB	Positive emotion ratio in class of beginning stage
IPM	Individual positive emotion ratio in middle stage
CPM	Positive emotion ratio in class of middle stage
IPS	Individual positive emotion ratio in summary stage
CPS	Positive emotion ratio in class of summary stage
INA	Individual negative emotion ratio
CNA	Negative emotion ratio in class
INB	Individual negative emotion ratio in beginning stage
CNB	Negative emotion ratio in class of beginning stage
INM	Individual negative emotion ratio in middle stage
CNM	Negative emotion ratio in class of middle stage
INS	Individual negative emotion ratio in summary stage
CNS	Negative emotion ratio in class of summary stage
ICA	Individual cognition ratio
CCA	Cognition ratio in class
ICB	Individual cognition ratio in beginning stage
CCB	Cognition ratio in class of beginning stage
ICM	Individual cognition ratio in middle stage
CCM	Cognition ratio in class of middle stage
ICS	Individual cognition ratio in summary stage
CCS	Cognition ratio in class of summary stage
IMA	Individual meta-cognition ratio
CMA	Meta-cognition ratio in class
IMB	Individual meta-cognition ratio in beginning stage
CMB	Meta-cognition ratio in class of beginning stage
IMM	Individual meta-cognition ratio in middle stage
CMM	Meta-cognition ratio in class of middle stage
IMS	Individual meta-cognition ratio in summary stage
CMS	Meta-cognition ratio in class of summary stage

(To be continued)

Table A1 Features set from the classroom dialogue dataset. (Continued)

Feature	Description
IOA	Individual off-topic ratio
COA	Off-topic ratio in class
IOB	Individual off-topic ratio in beginning stage
COB	Off-topic ratio in class of beginning stage
IOM	Individual off-topic ratio in middle stage
COM	Off-topic ratio in class of middle stage
IOS	Individual off-topic ratio in summary stage
COS	Off-topic ratio in class of summary stage
CTA	Text number ratio in class
ITB	Individual ratio of text number in beginning stage
CTB	Ratio of text number in class of beginning stage
ITM	Individual ratio of text number in middle stage
CTM	Ratio of text number in class of middle stage
ITS	Individual ratio of text number in summary stage
CTS	Ratio of text number in class of summary stage

Acknowledgment

This work was supported by the Center for Social Network Research of Tsinghua University, Tsinghua's Research Project (No. 2016THZWYY03), and the Project of Tencent Social Research Center (No. 20162001703).

References

[1] J. E. Knowles, Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin, *J. Educ. Data Min.*, vol. 7, no. 3, pp. 18–67, 2015.

[2] S. Lee and J. Y. Chung, The machine learning-based dropout early warning system for improving the performance of dropout prediction, *Appl. Sci.*, vol. 9, no. 15, p. 3093, 2019.

[3] L. Yan, A. Whitelock-Wainwright, Q. Guan, G. Wen, D. Gašević, and G. Chen, Students' experience of online learning during the COVID-19 pandemic: A province-wide survey study, *Br. J. Educ. Technol.*, vol. 52, no. 5, pp. 2038–2057, 2021.

[4] H. Li, W. Ding, and Z. Liu, Identifying at-risk K-12 students in multimodal online environments: A machine learning approach, arXiv preprint arXiv: 2003.09670, 2020.

[5] S. Hennessy, S. Rojas-Drummond, R. Higham, A. M. Márquez, F. Maine, R. M. Ríos, R. García-Carrión, O. Torreblanca, and M. J. Barrera, Developing a coding scheme for analysing classroom dialogue across educational contexts, *Learn. Cult. Soc. Interact.*, vol. 9, pp. 16–44, 2016.

[6] M. Zembylas, Adult learners' emotions in online learning, *Distance Educ.*, vol. 29, no. 1, pp. 71–87, 2008.

[7] Z. Liu, and Y. Zhang, A semantic role mining and learning performance prediction method in MOOCs, presented at Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Int. Conf. Web and Big Data, Macau, China, 2018, pp. 259–269.

[8] W. Zou, X. Hu, Z. Pan, C. Li, Y. Cai, and M. Liu, Exploring the relationship between social presence and learners' prestige in MOOC discussion forums using automated content analysis and social network analysis, *Comput. Hum. Behav.*, vol. 115, p. 106582, 2021.

[9] C. Howe and M. Abedin, Classroom dialogue: A systematic review across four decades of research, *Camb. J. Educ.*, vol. 43, no. 3, pp. 325–356, 2013.

[10] R. J. Alexander, *Essays on Pedagogy*. New York, NY, USA: Routledge, 2008.

[11] J. G. Greeno, Classroom talk sequences and learning, in *Socializing Intelligence Through Academic Talk and Dialogue*, L. B. Resnick, C. S. Asterhan, and S. N. Clarke, eds. New York, NY, USA: American Educational Research Association, 2015, pp. 255–262.

[12] H. Muhonen, E. Pakarinen, A. M. Poikkeus, M. K. Lerkkanen, and H. Rasku-Puttonen, Quality of educational dialogue and association with students' academic performance, *Learn. Instr.*, vol. 55, pp. 67–79, 2018.

[13] M. Baker, S. Järvelä, J. Andriessen, *Affective Learning Together: Social and Emotional Dimensions of Collaborative Learning*. New York, NY, USA: Routledge, 2013.

[14] G. Zhu, W. Xing, S. Costa, M. Scardamalia, and B. Pei, Exploring emotional and cognitive dynamics of Knowledge Building in grades 1 and 2, *User Model. User Adapt. Interact.*, vol. 29, no. 4, pp. 789–820, 2019.

[15] P. R. Kleinginna Jr and A. M. Kleinginna, A categorized list of emotion definitions, with suggestions for a consensual definition, *Motiv. Emot.*, vol. 5, no. 4, pp. 345–379, 1981.

[16] K. R. Scherer, What are emotions? And how can they be measured, *Soc. Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005.

[17] A. Bakhtiar, E. A. Webster, and A. F. Hadwin, Regulation and socio-emotional interactions in a positive and a negative group climate, *Metacognition Learn.*, vol. 13, no. 1, pp. 57–90, 2018.

[18] M. Cleveland-Innes and P. Campbell, Emotional presence, learning, and the online learning environment, *Int. Rev. Res. Open Distrib. Learn.*, vol. 13, no. 4, p. 269, 2012.

[19] A. R. Damasio, Descartes' error and the future of human life, *Sci. Am.*, vol. 271, no. 4, p. 144, 1994.

[20] R. Pekrun and L. Linnenbink-Garcia, Academic emotions and student engagement, in *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, and C. Wylie, eds. New York, NY, USA: Springer, 2012, pp. 259–282.

[21] L. Yorks and E. Kasl, Toward a theory and practice for whole-person learning: Reconceptualizing experience and the role of affect, *Adult Educ. Q.*, vol. 52, no. 3, pp. 176–192, 2002.

[22] R. Pekrun, A social cognitive, control–value theory of achievement emotions, in *Motivational Psychology of Human Development*, J. Heckhausen, ed. Oxford, UK: Elsevier, pp. 143–163, 2000.

[23] R. Pekrun, S. Lichtenfeld, H. W. Marsh, K. Murayama, and T. Goetz, Achievement emotions and academic performance: Longitudinal models of reciprocal effects, *Child Dev.*, vol. 88, no. 5, pp. 1653–1670, 2017.

[24] R. S. J. D. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A.

C. Graesser, Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments, *Int. J. Hum. Comput. Stud.*, vol. 68, no. 4, pp. 223–241, 2010.

[25] Z. A. Pardos, R. S. Baker, M. O. San Pedro, S. M. Gowda, and S. M. Gowda, Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes, in *Proc. 3rd Int. Conf. learning analytics and knowledge*, Leuven, Belgium, 2013, pp. 117–124.

[26] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser, Confusion can be beneficial for learning, *Learn. Instr.*, vol. 29, pp. 153–170, 2014.

[27] M. Worsley, P. Blikstein, Using learning analytics to study cognitive disequilibrium in a complex learning environment, in *Proc. 5th Int. Conf. Learning Analytics and Knowledge*, Poughkeepsie, NY, USA, 2015, pp. 426–427.

[28] L. Zheng, J. Niu, and L. Zhong, Effects of a learning analytics-based real-time feedback approach on knowledge elaboration, knowledge convergence, interactive relationships and group performance in CSCL, *Br. J. Educ. Technol.*, vol. 53, no. 1, pp. 130–149, 2022.

[29] L. Przybilla, K. Klinker, M. Kauschinger, and H. Krcmar, Stray off topic to stay on topic: Preserving interaction and team morale in a highly collaborative course while at a distance, *Commun. Assoc. Inf. Syst.*, vol. 48, no. 1, pp. 177–184, 2021.

[30] J. H. Flavell, Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry, *Am. Psychol.*, vol. 34, no. 10, pp. 906–911, 1979.

[31] M. E. Martinez, What is metacognition, *Phi Delta Kappan*, vol. 87, no. 9, pp. 696–699, 2006.

[32] C. D. Zepeda, C. O. Hlutkowsky, A. C. Partika, and T. J. Nokes-Malach, Identifying teachers' supports of metacognition through classroom talk and its relation to growth in conceptual learning, *J. Educ. Psychol.*, vol. 111, no. 3, pp. 522–541, 2019.

[33] P. H. Winne and J. C. Nesbit, The psychology of academic achievement, *Annu. Rev. Psychol.*, vol. 61, pp. 653–678, 2010.

[34] A. Cadamuro, E. Bisagno, G. A. Di Bernardo, L. Vezzali, and A. Versari, Making the school Smart: the interactive whiteboard against disparities in children stemming from low metacognitive skills, *J. E-Learn. Knowl. Soc.*, vol. 16, no. 1, pp. 33–43, 2020.

[35] C. M. Roebers, S. S. Krebs, and T. Roderer, Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance, *Learn. Individ. Differ.*, vol. 29, pp. 141–149, 2014.

[36] H. L. Swanson, Influence of metacognitive knowledge and aptitude on problem solving, *J. Educ. Psychol.*, vol. 82, no. 2, pp. 306–314, 1990.

[37] M. V. J. Veenman, R. Kok, and A. W. Blöte, The relation between intellectual and metacognitive skills in early adolescence, *Instr. Sci.*, vol. 33, no. 3, pp. 193–211, 2005.

[38] Z. R. Mevarech and C. Amrany, Immediate and delayed effects of meta-cognitive instruction on regulation of cognition and mathematics achievement, *Metacognition Learn.*, vol. 3, no. 2, pp. 147–157, 2008.

[39] R. J. Yinger, A study of teacher planning, *Elem. Sch. J.* , vol. 80, no. 3, pp. 107–127, 1980.

[40] M. Hunter, Knowing, teaching and supervising, in *Using What We Know About Reading*, P. Hosford, ed. Alexandria, Egypt: Association for Supervision and Curriculum Development, 1984, pp. 169–203.

[41] A. P. Johnson, It's time for madeline hunter to go: A new look at lesson plan design, *Action Teach. Educ.*, vol. 22, no. 1, pp. 72–78, 2000.

[42] T. S. Farrell, Lesson planning, in *Methodology in Language Teaching: An Anthology of Current Practice*, J. C. Richards and W. A. Renandya, Eds. New York, NY, USA: Cambridge University Press, 2002, pp. 30–39.

[43] G. Tonguç and B. O. Ozkara, Automatic recognition of student emotions from facial expressions during a lecture, *Comput. Educ.*, vol. 148, p. 103797, 2020.

[44] M. Ford, C. T. Baer, D. Xu, U. Yapanel, and S. Gray, The LENATM language environment analysis system: Audio specifications of the DLP-0121, Technical report, LENA Foundation, Boulder, CO, USA, 2008.

[45] Z. Wang, K. Miller, and K. Cortina, Using the LENA in teacher training: Promoting student involvement through automated feedback, *Unterrichtswissenschaft*, vol. 4, pp. 290–305, 2013.

[46] Z. Wang, X. Pan, K. F. Miller, and K. S. Cortina, Automatic classification of activities in classroom discourse, *Comput. Educ.*, vol. 78, pp. 115–123, 2014.

[47] D. Jiang, Y. Chen, and A. Garg, A hybrid method for overlapping speech detection in classroom environment, *Comput. Appl. Eng. Educ.*, vol. 26, no. 1, pp. 171–180, 2018.

[48] Y. Song, S. Lei, T. Hao, Z. Lan, and Y. Ding, Automatic classification of semantic content of classroom dialogue, *J. Educ. Comput. Res.*, vol. 59, no. 3, pp. 496–521, 2021.

[49] P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres, and C. Delgado Kloos, Analysing the predictive power for anticipating assignment grades in a massive open online course, *Behav. Inf. Technol.*, vol. 37, nos. 10–11, pp. 1021–1036, 2018.

[50] K. Mrhar, L. Benhiba, S. Bourekkache, and M. Abik, A Bayesian CNN-LSTM model for sentiment analysis in massive open online courses MOOCs, *Int. J. Emerg. Technol. Learn.* , vol. 16, no. 23, pp. 216–232, 2021.

[51] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *SSRN Electron. J.*, vol. 31, no. 2, pp. 841–887, 2017.

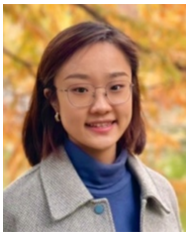
[52] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren, A survey of surveys on the use of visualization for interpreting machine learning models, *Inf. Vis.* , vol. 19, no. 3, pp. 207–233, 2020.

[53] G. Plumb, D. Molitor, A. S. Talwalkar, Model Agnostic Supervised Local Explanations, *NIPS*, vol. 31, pp. 2520–2529, 2018.

[54] M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.

[55] S. M. Lundberg, S. I. Lee, A unified approach to

interpreting model predictions, *NIPS*, vol. 30, pp. 4768–4777, 2017.
[56] S. Karlos, G. Kostopoulos, and S. Kotsiantis, Predicting and interpreting students' grades in distance higher education through a semi-regression method, *Appl. Sci.*, vol. 10, no. 23, p. 8413, 2020.
[57] S. Kim, W. Kim, Y. Jang, S. Choi, H. Jung, and H. Kim, Student knowledge prediction for teacher-student interaction, in *Proc. AAAI Conf. Artif. Intell.* , vol. 35, no. 17, pp. 15560–15568, 2021.
[58] J. Devlin, M. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805, 2018.
[59] J. Hoth, G. Kaiser, A. Busse, M. Doehrmann, J. Koenig, and S. Blömeke, Professional competences of teachers for fostering creativity and supporting high-achieving students, *ZDM Math. Educ.*, vol. 49, no. 1, pp. 107–120, 2017.
[60] J. T. Guthrie and A. Wigfield, Engagement and motivation in reading, in *Handbook of Reading Research*, P. D. Barr, ed. New York, NY, USA: Lawrence Erlbaum Associates, 2000, pp. 403–422.
[61] X. Xu, J. Wang, H. Peng, and R. Wu, Prediction of academic performance associated with Internet usage behaviors using machine learning algorithms, *Comput. Hum. Behav.*, vol. 98, pp. 166–173, 2019.
[62] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation, *Comput. Educ.*, vol. 2, p. 100018, 2021.
[63] L. Zheng, M. Long, J. Niu, and L. Zhong, An automated group learning engagement analysis and feedback approach to promoting collaborative knowledge building, group performance, and socially shared regulation in CSCL, *Int. J. Comput. Support. Collab. Learn.*, vol. 18, no. 1, pp. 101–133, 2023.



Yuanyi Zhen received the MS degree from Beijing Normal University, China in 2021. She is currently a PhD candidate at Department of Sociology, Tsinghua University, China. Her research interests include social computing and complex social theory.

Hui Chen is currently an associate professor at School of Chinese Language and Literature, Beijing Foreign Studies University, China. Her research interests include natural language process and educational technology.

Jar-Der Luo received the PhD degree from State University of New York at Stony Brook, USA in 1993. He is currently a professor at Department of Sociology, Tsinghua University, president of Chinese Network for Social Network Studies, and chairman of Tsinghua Social Network Research Center, China. His research interests focus on social capital, trust, social network analysis in big data, self-organization process, and Chinese indigenous management researches, such as guanxi and guanxi circle.