Computers and Education: Artificial Intelligence 8 (2025) 100366

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Perceived MOOC satisfaction: A review mining approach using machine learning and fine-tuned BERTs

Xieling Chen [a], Haoran Xie [b,*], Di Zou [c], Gary Cheng [d], Xiaohui Tao [e], Fu Lee Wang [f]

[a] School of Education, Guangzhou University, China
[b] School of Data Science, Lingnan University, Hong Kong
[c] Department of English and Communication, The Hong Kong Polytechnic University, Hong Kong
[d] Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong
[e] School of Mathematics, Physics and Computing, University of Southern Queensland, Australia
[f] School of Science and Technology, Hong Kong Metropolitan University, Hong Kong

ARTICLE INFO

Keywords:
MOOCs
Learner satisfaction
Machine learning
BERT models
Multiple linear regression

ABSTRACT

This study investigates the application of machine learning and BERT models to identify topic categories in helpful online course reviews and uncover factors that influence the overall satisfaction of learners in massive open online courses (MOOCs). The research has three main objectives: (1) to assess the effectiveness of machine learning models in classifying review helpfulness, (2) to evaluate the performance of fine-tuned BERT models in identifying review topics, and (3) to explore the factors that influence learner satisfaction across various disciplines. The study uses a MOOC corpus containing 102,184 course reviews from 401 courses across 13 disciplines. The methodology involves three approaches: (1) machine learning for automatic classification of review helpfulness, (2) BERT models for automatic classification of review topics, and (3) multiple linear regression analysis to explore the factors influencing learner satisfaction. The results show that most machine learning models achieve precision, recall, and F1 scores above 80%, 99%, and 89%, respectively, in identifying review helpfulness. The fine-tuned BERT model outperforms baseline models with precision, recall, and F1 scores of 78.4%, 74.4%, and 75.9%, respectively, in classifying review topics. Additionally, the regression analysis identifies key factors affecting learner satisfaction, such as the positive influence of “Instructor” frequency and the negative impact of “Platforms and tools” and “Process”. These insights offer valuable guidance for educators, course designers, and platform developers, contributing to the optimization of MOOC offerings to better meet the evolving needs of learners.

1. Introduction

Massive open online courses (MOOCs), as a widely adopted format of digital education, have drastically changed the nature of education in the digital age by providing distance learners with abundant learning materials and interactive environments (Bitakou et al., 2023; Fang et al., 2022; Liu et al., 2023). A growing number of educational institutions have developed MOOCs, offering learners opportunities to share perspectives on their learning experiences by writing course reviews. These reviews contribute to a large-scale body of learner-generated content that is available for educational data mining (Chen et al., 2024) and provides insights into learning experiences (Hew et al., 2020; Ranga et al., 2023; Zhuang et al., 2023). Frequently containing valuable commentary, these reviews form a substantial informational base that can guide teachers, course designers, and prospective students. Efficiently navigating through a wide range of viewpoints requires the application of sophisticated analytical methods to assess the value of reviews and identify the underlying topics that students find compelling.

The well-established influence of online reviews extends to online education. Reviews play a crucial role in facilitating the decision-making process for learners (Chakraborty & Biswal, 2023). However, few studies have focused on the helpfulness of MOOC reviews, highlighting the need for further investigation. While previous studies have employed machine learning classifiers to categorize topics in MOOC reviews, the emergence and applicability of deep learning methods, especially bidirectional encoder representations from transformers

* Corresponding author.
E-mail addresses: xielingchen0708@gmail.com (X. Chen), hrxie2@gmail.com (H. Xie), dizoudaisy@gmail.com (D. Zou), chengks@eduhk.hk (G. Cheng), Xiaohui.Tao@usq.edu.au (X. Tao), pwang@hkmu.edu.hk (F. Lee Wang).

https://doi.org/10.1016/j.caeai.2025.100366
Received 2 November 2024; Received in revised form 7 January 2025; Accepted 15 January 2025
Available online 17 January 2025

(BERT) models, have garnered much attention in recent years (Li et al., 2020). BERT models demonstrate exceptional proficiency in understanding context and semantic nuances, making them well-suited for extracting meaningful topics from textual data.

As BERT's effectiveness is widely demonstrated in addressing different natural language processing (NLP) tasks (e.g., Kaur & Kaur, 2023; Suzuki et al., 2023; Wang et al., 2024), prompting researchers to apply it in advancing intelligent education applications. For example, Wulff et al. (2023) used BERT to classify segments of preservice physics teachers' reflections, while Cavalcanti et al. (2020) employed BERT to classify Portuguese instructors' feedback texts. In MOOCs, Sebbaq and El Faddouli (2022) used BERT-based transfer learning for automatic MOOC pedagogical annotation, focusing on students' cognitive levels. Despite these notable applications in NLP tasks, the use of BERT models to classify course review topics within MOOC learning contexts remains relatively limited.

Utilizing review data can provide a less intrusive method for collecting and storing students' learning data (Chen et al., 2024) compared to interviews or questionnaires, thus enhancing data quality and usability. As learning analytics progresses, these abundant datasets are increasingly analyzed to gain insights into students' experiences and the success of MOOCs in engaging learners, ultimately informing instructional design. Although existing research has identified factors influencing MOOC learners' satisfaction, there is an urgent need to explore subject-specific nuances. The dynamic nature of online learning environments necessitates ongoing exploration of factors influencing overall satisfaction, both holistically and within specific subject domains.

2. Research objectives and questions

This study investigates the effectiveness of machine learning and BERT models in online course review classification and in understanding the factors that affect MOOC learner satisfaction. Specifically, the study has three objectives:

Objective 1: Assess the potential of machine learning for classifying review helpfulness.

Objective 2: Assess the potential of BERT models for classifying review topics.

Objective 3: Identify the factors contributing to the overall satisfaction of MOOC learners across different subject domains.

Accordingly, there are three main research questions (RQs).

RQ1: To what extent can machine learning models effectively discern helpful online course reviews?

RQ2: To what extent can BERT models accurately identify the topics addressed in online course reviews?

RQ3: What factors contribute to MOOC learner satisfaction, both collectively and within distinct subject domains?

This study addresses RQ1 by examining various machine learning algorithms to identify the optimal and most broadly applicable model for automating the classification of MOOC reviews, particularly with regard to perceived helpfulness across diverse subject areas. RQ2 is addressed by leveraging and comparing BERT's capabilities with baseline models to improve the accuracy of online course review topic classification. To address RQ3, the study adopts a granular approach, segmenting data by subject area to provide a nuanced understanding of the various factors contributing to learner satisfaction. This approach bridges the gap between general satisfaction trends and subject-specific factors within the context of online education.

3. Literature review

3.1. Machine learning in identifying helpful reviews

Online feedback consists of favorable, indifferent, or adverse evaluations regarding a company, its products, or its services. Typically, these evaluations are accompanied by a numerical rating and are posted on digital platforms by individuals who claim to have experienced or purchased the reviewed item or service. However, not all comments and/or rating scores in reviews related to specific products hold equal significance for evaluating product performance. Certain review comments may not be considered as informative as others by consumers seeking to familiarize themselves with a product or service and assess its quality (Ganguly et al., 2024). Consequently, an increasing number of scholars are focusing their investigations on identifying review helpfulness (e.g., Kong & Lou, 2023; Zhou et al., 2023). For example, Quaderi and Varathan (2024) examined the potential of features such as linguistics, metadata, readability, subjectivity, and polarity for predicting online review helpfulness based on machine learning.

Previous research has primarily focused on course review helpfulness in the domains of restaurants, hotels, and tourism. In contrast, the consideration of review helpfulness in online learning has been largely overlooked (Meek et al., 2021). The online learning field operates similarly to online commerce, where the quality of a review is determined by the helpfulness of its information as judged by reader votes. Notably, both business reviews and MOOC course reviews share similar characteristics. Hence, when conducting research using online course review data, it becomes imperative to consider review helpfulness, particularly through the use of artificial intelligence-driven educational big data and learning analytics, which have been identified as key technological components in smart learning environments (Darmawansah et al., 2023; Hwang & Fu, 2020). For example, Lubis et al. (2017) introduced program prototypes for automatically classifying helpful MOOC reviews using Naive Bayes and sentiment analysis. Lubis et al. (2019) applied Naive Bayes to classify review helpfulness based on course reviews collected from Class Central.

Existing studies have typically employed a single machine learning algorithm, focusing solely on review data related to programming courses. This limited scope may result in a classifier that is less broadly applicable. Therefore, it is crucial to consider and compare multiple machine learning algorithms to identify an optimal, more universally applicable model for the automated classification of MOOC reviews across various subject domains based on their perceived helpfulness. Additionally, while using "Like" or "helpful votes" as a metric for review helpfulness is common, it typically requires time for a review to accumulate "Likes". In other words, reviews that have been posted for a longer period are more likely to receive "Likes". For reviews posted recently, there is a possibility that no "Likes" were received at the time they were collected.

Given the evidence supporting the efficacy of machine learning in classifying review helpfulness, there is a need to automatically recognize helpful reviews among newly submitted ones. This is crucial for understanding learners' perceptions of course content based on up-to-date and relevant reviews. By thoroughly investigating the effectiveness of various machine learning approaches, this study aims to enhance and optimize models designed to recognize the nuanced factors influencing review helpfulness.

3.2. Automatic classification of online course review topics

Classification is a common challenge in various fields, including education, and addressing classification tasks using supervised learning approaches is crucial for achieving intelligent classification in the era of big data (Zhou et al., 2023). Text classification is a traditional challenge within NLP, aiming to assign labels or tags to textual components such as sentences, paragraphs, and documents (Soni et al., 2023). In online education, scholars have employed NLP techniques to analyze MOOC review data to understand the correlation between learners' activities in forums and their overall success. Recently, deep learning has proven effective in various MOOC-related tasks by incrementally learning high-level features from data. For instance, Liu et al. (2024) introduced deep learning networks that integrate learning behavior features to predict dropout rates in MOOCs. Wen and Juan (2024) presented a deep learning network model that integrated learning behavior features to predict dropout rates in MOOCs. Koufakou (2024) employed machine learning techniques, including word embeddings and advanced neural networks like BERT, RoBERTa, and XLNet, to analyze student sentiments and topics effectively. Chen et al. (2022) utilized sentiment analysis and deep neural networks to examine MOOC students' reviews, aiming to identify key variables associated with course design and understand students' learning perceptions. Gupta et al. (2023) proposed a deep learning methodology that uses facial expressions to identify the immediate engagement of online learners by analyzing their facial reactions throughout the entire online learning session.

BERT, a transformer-based approach for natural language understanding (Xue et al., 2023), has demonstrated excellence across various language-related tasks, extending beyond written reflections (Wulff et al., 2023). In online education, for example, Srivastav et al. (2024) introduced an automated feedback assessment model that employs BERT to generate quality scores for inputs in virtual learning environments. Zhu et al. (2022) employed a fine-tuned BERT framework to automatically encode answer texts, thereby enhancing Short-Answer Grading. However, since BERT's release, there have been few studies exploring its efficacy in evaluating the topics of online course reviews. Therefore, this study utilizes BERT to improve the performance of topic classification in online course reviews.

3.3. *Research on influencing factors for MOOC satisfaction*

The exploration of factors influencing the overall satisfaction of MOOC learners remains a prominent theme in educational research. A growing body of studies has examined the determinants of satisfaction among MOOC learners. Initially, researchers systematically reviewed scientific papers and survey studies to uncover these factors. For example, Albelbisi (2020) reviewed relevant studies on MOOC success measurement and identified factors such as "system quality", "information quality", "service quality", "attitude", and "course quality" that could impact learner satisfaction. Based on the expectation confirmation model, Alraimi et al. (2015) highlighted the significant effects of factors such as "perceived usefulness", "enjoyment", and "support" on satisfaction. Drawing on self-determination theory, Pozón-López et al. (2021) demonstrated the significant effects of learner-perceived satisfaction and autonomous motivation on continuance intention.

Researchers have also employed machine learning and statistical analysis approaches to understand MOOC review topics and their impact on student satisfaction. For instance, Alsayat and Ahmadi (2023) conducted an analysis of learner satisfaction using text mining and supervised learning techniques, employing ensemble learning methods. They used the AdaBoost boosting technique within artificial neural networks to enhance its performance. Sandiwarno et al. (2024) introduced a machine learning model that combined features extracted from user activities, usability testing, and user opinions, using a convolutional neural network and bidirectional long short-term memory. Chen et al. (2020) applied structural topic modeling to analyze 1920 reviews from learners enrolled in 339 computer science MOOCs, identifying factors such as "course levels", "learning perception", "course assessment", "teaching styles", "problem-solving", "course content", "course organization", "critique", and "learning tools and platforms" that were of significant concern to learners. Liu et al. (2019) used an optimized behavior-sentiment topic mixture model to automatically analyze learners' opinions about courses, instructors, and MOOC systems, revealing that "course-related content", "course logistics", and "video production" were important factors influencing MOOC learner satisfaction. However, the dynamic nature of online education, coupled with students' evolving expectations, underscores the need for continuous exploration of the factors influencing overall satisfaction, both broadly and within specific subject domains.

This study aims not only to conduct a comprehensive evaluation of MOOC satisfaction but also to facilitate a detailed examination within specific subject domains. By segmenting the data based on subject areas, the study seeks to determine whether certain factors exert a more significant influence in particular domains, offering a nuanced understanding of the various elements contributing to learner satisfaction. Ultimately, such analysis will provide a holistic view of the complex factors shaping MOOC learners' satisfaction, bridging the gap between overarching satisfaction trends and subject-specific intricacies within the vast landscape of online education.

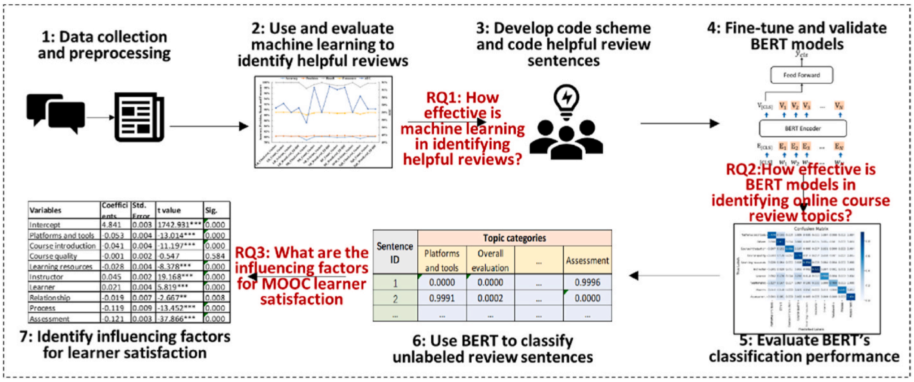
4. Research methodology

This study employs three types of methods to address the three RQs, including (1) machine learning for the automatic classification of MOOC review helpfulness, (2) BERT models for the automatic classification of MOOC review topics, and (3) multiple linear regression analysis to explore factors that influence MOOC learner satisfaction, both collectively and within distinct subject domains. Fig. 1 illustrates the research design.

Step 1: The original dataset, which includes extensive course review data and course metadata, was collected from the Class Central platform. The data was pre-processed using NLP tools to construct a MOOC corpus containing 102,184 reviews.
Step 2: This step addresses RQ1. It involves the task of automatically classifying MOOC review helpfulness using machine learning. Of the 102,184 reviews, 8,517 reviews with helpful votes were categorized as either helpful (6,813) or unhelpful (1,704) based on the number of helpful votes. These were randomly assigned to training (80%) and testing (20%) datasets to train and test machine learning models. The models were then used to automatically classify 93,667 reviews (102,184–8,517) without helpful votes, identifying an additional 92,966 helpful reviews.
Step 3: In the task of automatically classifying MOOC review topics using BERT models, 99,779 helpful reviews (6,813 + 92,966) were selected, which together contained a total of 402,188 review sentences. From these, 11,732 review sentences were randomly selected and coded according to their associated topic categories based on a coding scheme developed by domain experts and coders for analyzing assessments within MOOC participant reviews.
Step 4: This step addresses RQ2. The BERT model was evaluated against six baseline approaches using the 11,732 coded helpful review sentences. The classification performance of the models was assessed using precision, recall, and F1-score.
Step 5: The BERT model's performance across different topic categories was further evaluated using precision, recall, and F1-score. Visualization was achieved by plotting confusion matrices.
Step 6: The BERT model was used to automatically categorize 390,456 review sentences with topic category labels.
Step 7: This step addresses RQ3. Multiple linear regression analysis was implemented to explore the factors influencing the satisfaction of MOOC learners across different discipline domains. This was based on the frequency of topic categories identified from the 402,188 review sentences (390,456 + 11,732) and learners' overall satisfaction levels, indicated by an overall rating score given by the learners.

6.1. *Data collection and preprocessing*

The MOOC reviews were gathered from Class Central. After excluding duplicates and MOOCs with fewer than 20 reviews, non-English reviews were eliminated using the "langid" Python package. The "langid" is a fast and efficient Python package used for language identification. It is a standalone tool that can automatically detect the


1: Data collection and preprocessing
2: Use and evaluate machine learning to identify helpful reviews
3: Develop code scheme and code helpful review sentences
4: Fine-tune and validate BERT models
Feed Forward
BERT Encoder
RQ1: How effective is machine learning in identifying helpful reviews?
RQ2:How effective is BERT models in identifying online course review topics?
RQ3: What are the influencing factors for MOOC learner satisfaction
5: Evaluate BERT's classification performance
6: Use BERT to classify unlabeled review sentences
7: Identify influencing factors for learner satisfaction

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.841	0.003	1742.931***	0.000
Platforms and tools	-0.053	0.004	-13.014***	0.000
Course introduction	-0.041	0.004	-11.197***	0.000
Course quality	-0.001	0.002	-0.547	0.584
Learning resources	-0.028	0.004	-8.378***	0.000
Instructor	0.045	0.002	19.168***	0.000
Learner	0.021	0.004	5.819***	0.000
Relationship	-0.019	0.007	-2.667**	0.008
Process	-0.119	0.009	-13.452***	0.000
Assessment	-0.121	0.003	-37.866***	0.000

Sentence ID	Topic categories			
	Platforms and tools	Overall evaluation	...	Assessment
1	0.0000	0.0000	...	0.9996
2	0.9991	0.0002	...	0.0000
...	...	...	...	...

Fig. 1. Research design.

language of a given text. The package employs machine learning models trained on a large corpus of textual data to classify input text into one of 97 supported languages. In our study, we used the “langid” to identify reviews written in languages other than English and excluded them from our dataset. This ensured that our analysis focused solely on English-language reviews, which is crucial for the consistency and quality of our research.

After removing learner privacy information from all data, the texts were segmented into word units, spellings were corrected, and stop words were excluded using the Natural Language Toolkit. The Python tool TextBlob was then employed to automatically check and correct spelling. Ultimately, a MOOC corpus was compiled, encompassing 102,184 course reviews from 401 courses across 13 discipline domains. This dataset is significantly larger than those used in previous studies with similar objectives. For example, the dataset used in a study on automatic MOOC review topic classification (Hew et al., 2020) contained only 6393 reviews. Similarly, for the task of automatic MOOC review helpfulness classification, Lubis et al. (2019) used around 5000 reviews from Class Central, and Lubis et al. (2017) utilized 5340 reviews as raw data. By contrast, our dataset, with 102,184 reviews, offers substantially more data for model training and evaluation, which we believe enhances the generalizability and robustness of our models across various disciplines and topic categories. Additionally, the dataset spans 13 discipline domains, further ensuring diverse and comprehensive coverage. Given the scale of the dataset, we are confident in its adequacy for the development of an optimal and adaptable model for classifying both review helpfulness and topic categories.

6.2. Helpful review identification using machine learning

This study addresses the issue of some reviews lacking useful information about specific aspects of MOOCs by training and testing machine learning classifiers for the automatic identification of review helpfulness. Fig. 2 illustrates the step-by-step approaches for review helpfulness classification using machine learning. There are five steps: identifying reviews with helpful votes, defining helpful and unhelpful reviews, transforming raw data into feature vectors, constructing classification models, and evaluating classification performance. The details are described as follows.

6.2.1. Identifying reviews with helpful votes

Out of the 102,184 reviews, only 8,517 received helpful votes from other learners. Helpful votes serve as a peer-assessed indicator of the quality, relevance, and informativeness of a review (Liu & Park, 2015). By focusing on reviews with helpful votes, we ensure that our analysis is grounded in content that has been validated as meaningful and representative of learner experiences (O’Mahony & Smyth, 2010). This approach aligns with methodologies in previous research (e.g., Chen et al., 2020; Pozón-López et al., 2021), where only reviews with helpful votes or other quality indicators are included to ensure reliable and

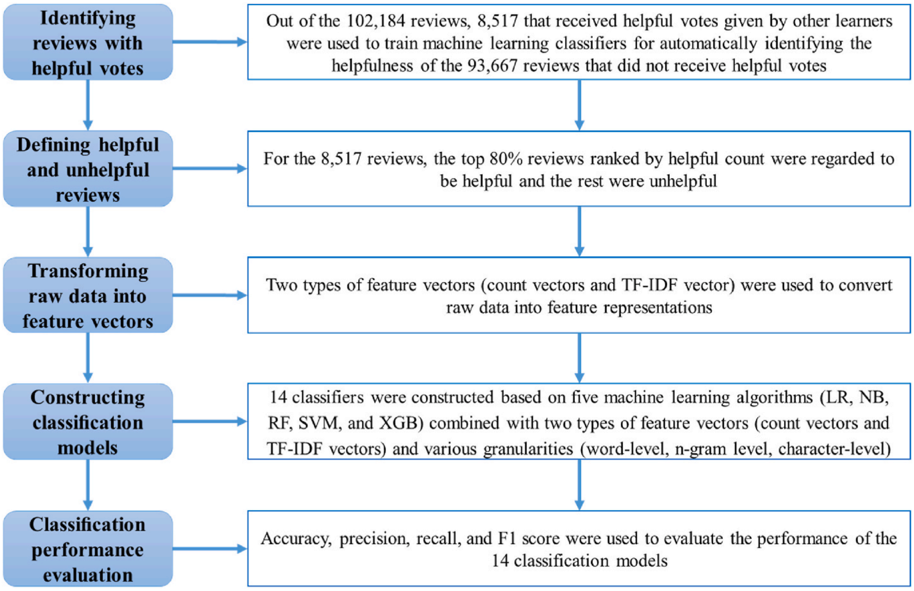

Identifying reviews with helpful votes
Out of the 102,184 reviews, 8,517 that received helpful votes given by other learners were used to train machine learning classifiers for automatically identifying the helpfulness of the 93,667 reviews that did not receive helpful votes
Defining helpful and unhelpful reviews
For the 8,517 reviews, the top 80% reviews ranked by helpful count were regarded to be helpful and the rest were unhelpful
Transforming raw data into feature vectors
Two types of feature vectors (count vectors and TF-IDF vector) were used to convert raw data into feature representations
Constructing classification models
14 classifiers were constructed based on five machine learning algorithms (LR, NB, RF, SVM, and XGB) combined with two types of feature vectors (count vectors and TF-IDF vectors) and various granularities (word-level, n-gram level, character-level)
Classification performance evaluation
Accuracy, precision, recall, and F1 score were used to evaluate the performance of the 14 classification models

Fig. 2. Step-by-step approaches of review helpfulness classification using machine learning.

insightful analysis. Following this established practice enhances the comparability and credibility of our findings within the research community. For the 93,667 reviews that did not receive helpful votes, we did not exclude them directly in the analysis of influential factors on MOOC learner satisfaction. This is because reviews posted earlier are more likely to have received helpful votes due to their longer visibility, while more recently posted reviews may not have received sufficient exposure to accumulate helpful votes. Therefore, we trained a machine learning classifier using the 8,517 reviews that received helpful votes to automatically identify the helpfulness of the 93,667 reviews.

6.2.2. *Defining helpful and unhelpful review*

To define a helpful review, a common practice in the literature is to use the helpful vote ratio (e.g., Lee et al., 2018; O'Mahony & Smyth, 2009). This study followed O'Mahony and Smyth (2009)'s (2009) methodology, ranking the 8,517 reviews in descending order based on the number of helpful votes they received and adopting a top-percentile approach. The top 80% of reviews (6,813 reviews) were categorized as helpful, while the remaining 20% (1,704 reviews) were categorized as unhelpful. This threshold was chosen to focus our analysis on the reviews most consistently recognized as helpful by MOOC users, ensuring a clearer distinction between highly impactful reviews and less impactful ones. This approach is appropriate and aligns with established practices in the field, providing a data-driven and standardized way to distinguish between levels of helpfulness while maintaining analytical focus. Although the 1,704 reviews categorized as unhelpful did receive some helpful votes, their ranking in the bottom 20% indicates that they were less consistently deemed helpful compared to the top 80%. By using this ranking-based method, we ensure that our analysis emphasizes reviews with the highest perceived value and relevance, as determined by user voting patterns. The 8,517 reviews were randomly assigned to training (80% of the dataset) and testing (20%) datasets to compare the performance of different machine learning models in review helpfulness classification.

6.2.3. *Transforming raw data into feature vectors*

This study used two different types of feature vectors to convert raw data into feature representations: count vectors and TF-IDF vectors. Count vectors represent the frequency of terms (words) in the reviews, with each element corresponding to the count of a specific word in a review. TF-IDF vectors capture term importance by giving more weight to terms that are frequent in specific reviews but rare across the corpus, based on input tokens such as words, characters, and N-grams. In Word-Level TF-IDF, the matrix refers to the TF-IDF scores of each term in the texts. In N-gram-Level TF-IDF, the matrix reflects the TF-IDF scores for N-grams. In Character-Level TF-IDF, the matrix denotes the TF-IDF scores for character-level N-grams in the corpus.

6.2.4. *Construction classification models*

This study applied five machine learning algorithms for review helpfulness classification: logistic regression (LR), naive Bayes (NB), random forest (RF), support vector machine (SVM), and XGBoost. These algorithms were combined with two different types of feature vectors (count vectors and TF-IDF vectors) and three granularities (word level, n-gram level, and character level), resulting in a total of 14 classifiers. The details of these 14 classifiers are as follows.

(1) LR with Count Vectors (LR_Count_Vectors) classifier: uses LR as the machine learning algorithm and employs count vectors to convert raw data into feature representations
(2) LR with Word Level TF-IDF (LR_WordLevel_TF-IDF) classifier: uses LR as the machine learning algorithm and employs word-level TF-IDF vectors to convert raw data into feature representations
(3) LR with N-gram Level TF-IDF (LR_N-Gram_Vectors) classifier: uses LR as the machine learning algorithm and employs N-gram level TF-IDF vectors to convert raw data into feature representations
(4) LR with Character Level TF-IDF (LR_CharLevel_Vectors) classifier: uses LR as the machine learning algorithm and employs character-level TF-IDF vectors to convert raw data into feature representations
(5) NB with Count Vectors (NB_Count_Vectors) classifier: uses NB as the machine learning algorithm and employs count vectors to convert raw data into feature representations
(6) NB with Word Level TF-IDF (NB_WordLevel_TF-IDF) classifier: uses NB as the machine learning algorithm and employs word-level TF-IDF vectors to convert raw data into feature representations
(7) NB with N-gram Level TF-IDF (NB_N-Gram_Vectors) classifier: uses NB as the machine learning algorithm and employs N-gram level TF-IDF vectors to convert raw data into feature representations
(8) NB with Character Level TF-IDF (NB_CharLevel_Vectors) classifier: uses NB as the machine learning algorithm and employs character-level TF-IDF vectors to convert raw data into feature representations
(9) RF with Count Vectors (RF_Count_Vectors) classifier: uses RF as the machine learning algorithm and employs count vectors to convert raw data into feature representations
(10) RF with Word Level TF-IDF (RF_WordLevel_TF-IDF) classifier: uses RF as the machine learning algorithm and employs word-level TF-IDF vectors to convert raw data into feature representations
(11) SVM with N-gram Level TF-IDF (SVM_N-Gram_Vectors) classifier: uses SVM as the machine learning algorithm and employs N-gram level TF-IDF vectors to convert raw data into feature representations
(12) XGBoost with Character Level TF-IDF (Xgb_CharLevel_Vectors) classifier: uses XGB as the machine learning algorithm and employs character-level TF-IDF vectors to convert raw data into feature representations
(13) XGBoost with Count Vectors (Xgb_Count_Vectors) classifier: uses XGB as the machine learning algorithm and employs count vectors to convert raw data into feature representations
(14) XGBoost with Word Level TF-IDF (Xgb_WordLevel_TF-IDF) classifier: uses XGB as the machine learning algorithm and employs word-level TF-IDF vectors to convert raw data into feature representations

6.2.5. *Classification performance evaluation*

Using different combinations of these algorithms and feature vectors, a total of 14 classifiers were trained and tested to compare their performance in identifying helpful course reviews based on accuracy, precision, recall, and F1 score. Accuracy is the proportion of correct predictions among all cases. Precision reflects the proportion of correct predictions among predicted positive cases. Recall suggests the proportion of true positives among all actual positive cases. The F1 score is the harmonic mean of precision and recall.

6.3. *Coding analysis*

Drawing on existing design models and classifications found in the literature, which are used for analyzing assessments within MOOC participant reviews (e.g., Chen et al., 2022, 2024), a coding scheme is proposed. This scheme comprises ten categories, as outlined in Table A1 in the **Appendix**, facilitating the annotation of an "instructional" dataset. This dataset is intended for classifier training, validation, and testing, with the capability to classify the topics referred to in a review.

The "Course introduction" category encompasses descriptions of course information, such as the syllabus, schedule/calendar, certificates, and instructional languages. The "Course quality" category addresses content quality, information quality, course difficulty, course usefulness, and considerations of whether the course is beginner-friendly or enhances knowledge. The “Learning resources” category focuses on the availability of learning materials, including textbooks, notes, handouts, slides, and additional links. The “Instructor” category evaluates course instructors based on factors like instructor knowledge, enthusiasm, and humor. The “Learner” category encompasses information about students’ backgrounds, interests, and needs. The “Relationship” category evaluates the level of interaction, while the “Assessment” category assesses the quality of quizzes, assignments, projects, exercises, and tests. The “Process” category evaluates aspects such as feedback provision, learning activities, problem-solving, and the use of examples. The “Platforms and tools” category concentrates on the use of platforms, system quality, and video quality. Finally, the “Others” category captures students’ overall perceptions or attitudes.

To explore the application of BERT in categorizing review topics, 11,732 review sentences were randomly selected from the 402,188 review sentences that comprise the full dataset of 99,779 helpful reviews. The selected review sentences were coded by two experienced coders following the established coding scheme, creating an evaluation dataset containing relevant comments on review topics. A Cohen’s kappa coefficient of 92.4% indicates sufficient reliability between the coders. Additionally, another experienced MOOC expert meticulously reviewed the codes and resolved any inconsistencies.

6.4. Identifying review topics using BERT models

In this study, we employed BERT, a state-of-the-art language model, for the task of identifying topics in course reviews. BERT is pre-trained on large-scale texts and has demonstrated exceptional performance across a wide range of NLP tasks. To make BERT more effective for our specific task of course review topic identification, we fine-tuned the model. Fine-tuning adjusts BERT’s parameters to better understand the specific context and language of course reviews, which are domain-specific (Wang et al., 2024).

Fig. 3 shows the architecture of the BERT model, fine-tuned for the task of course review topic identification. The model consists of three components: the input layer, the BERT encoder layer, and the output layer. Specifically, the input layer processes the raw text from a course review. The input begins with the special classification token $[CLS]$, followed by a sequence of tokens $[w_1, w_2, ., w_N]$ derived from the text. Each token is transformed into an embedding $[E_1, E_2, ., E_N]$, serving as input to the BERT encoder layer, which applies a series of transformer blocks to the input embeddings to capture the contextual relationships between tokens. The transformer mechanism enables BERT to understand the meaning of each word in the context of the entire input sequence. This process generates a set of contextualized embeddings $[V_{[CLS]}, V_1, V_2, ., V_N]$, where $[V_{[CLS]}$ aggregates information from the entire sequence and is used for classification tasks. The contextualized embedding $[V_{[CLS]}$, produced by the BERT encoder layer, is passed through a fully connected feed-forward neural network in the output layer. The output layer uses a Softmax activation function to compute the probability distribution of categories and determine the predicted category with the highest probability. For example, in Fig. 3, the model assigns a probability of 0.9996 to the “Assessment” category, indicating that this category is the most likely topic of the given review.

By fine-tuning BERT on the course review dataset, the model learns task-specific patterns, enabling it to classify course reviews into predefined topic categories with high accuracy. The parameters of the model are updated during training via the backpropagation process, ensuring that the model is optimized for the task of review topic identification.

To evaluate BERT’s performance, we compared it against six baseline methods commonly used for text classification. Each of these models leverages Word2Vec embeddings to represent words as numerical vectors, but they apply different types of neural networks to capture and process relationships and patterns in the text. The six baseline methods are as follows.

(1) Word2vec + CNN: Word2Vec is used to represent the text (course reviews) as sequences of word vectors, and a convolutional neural network (CNN) is applied on top of those vectors to learn spatial hierarchies in the data, which is useful for identifying topics in text
(2) Word2vec + FastText: FastText improves Word2Vec by incorporating information about subword structures, which is beneficial for text classification, especially when the vocabulary includes many rare or unseen words.
(3) Word2vec + CRNN: Word2Vec vectors are passed through CNN layers to extract local features (e.g., n-grams or patterns in sequences), and then the output of the CNN is passed through recurrent neural network (RNN) layers to capture the temporal or sequential relationships in the review text.
(4) Word2vec + HAN: Word2Vec vectors are used to represent the words in a review, and hierarchical attention networks (HAN) then processes these vectors hierarchically, learning the importance of words within sentences and the importance of sentences within the entire review. The attention mechanism allows the model to focus on the most relevant information for classification, improving topic identification.

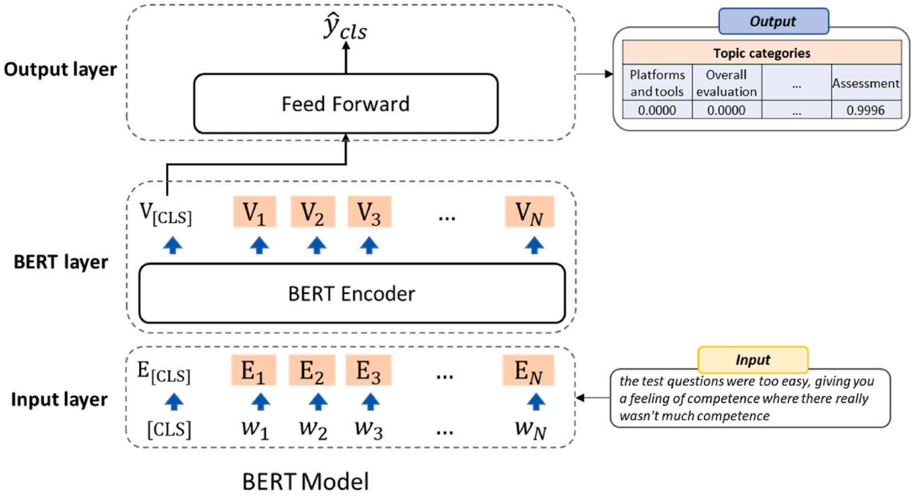


Fig. 3. Architecture of the BERT model.

(5) Word2vec + RNN: Word2Vec vectors are fed into an RNN to capture the sequential structure of the text.
(6) Word2vec + SANN: Word2Vec embeddings are passed through a self-attention mechanism, which learns which words in the review are most important for identifying the review's topic.

All experiments were executed on a single NVIDIA RTX 3080 (16 GB) GPU. For Word2Vec-based baselines, a batch size of 256 and a dropout rate of 0.5 were used to train the classifier with cross-entropy loss. The Adam optimizer was employed with an initial learning rate ranging from 1e-2 to 1e-5, and the model with the best performance was saved. For the BERT models (Devlin et al., 2018), a batch size of 8 and an initial learning rate of 3e-5, determined through grid search, were used. In all models, the hidden size for recurrent modules was set to 256, and the number of kernels for convolutional modules was set to 256.

The evaluation and comparison of BERT's classification performance, both before and after fine-tuning, along with the six baseline techniques, were conducted using three metrics: precision, recall, and F1-score. These metrics, commonly used in text categorization, provide insights into the effectiveness of classification models.

6.5. *Identifying influencing factors for MOOC learners' satisfaction using multiple linear regression analysis*

For RQ3, multiple linear regression analysis was applied to explore the factors contributing to MOOC learner satisfaction, both collectively and across distinct subject domains. Specifically, the factors analyzed include the frequency of topic categories identified from 402,188 review sentences (comprising 390,456 auto-classified and 11,732 manually coded sentences), which served as the independent variables in the regression model. These sentences were extracted from a total of 99,779 helpful reviews. More specifically, 11,732 review sentences were randomly selected from the dataset of 99,779 helpful reviews (6,813 + 92,966). These sentences were manually coded according to topic categories. The remaining 390,456 review sentences were automatically categorized into topic categories using the BERT classifier, which was trained on the manually coded dataset. The resulting topic categories were then used in the multiple linear regression analysis to examine their relationship with learner satisfaction, as measured by the learners' overall ratings.

By considering the learners' overall ratings as the dependent variable in the regression analysis, this study aimed to reveal the nuanced interplay between the frequency of topic categories in review comments and learners' overall satisfaction levels. In this context, a variable demonstrating a significantly positive effect on learner satisfaction implies that learners are content with that particular aspect. Conversely, a variable exhibiting a significantly negative effect suggests dissatisfaction among learners regarding that aspect.

This analysis facilitated both a comprehensive evaluation of MOOC satisfaction and a detailed examination within specific subject domains. By dividing the data according to subject areas, we were able to identify whether certain factors had a stronger influence in particular domains, thus gaining a more nuanced understanding of the elements contributing to learner satisfaction. Essentially, this multifaceted approach to multiple linear regression analysis offered a thorough perspective on the complex factors affecting MOOC learner satisfaction, bridging the gap between overarching satisfaction patterns and subject-specific intricacies within the vast landscape of online education.

5. Results

5.1. *Helpful online course reviews identification by machine learning*

Table 1
Evaluation of the 14 classifiers.

Classifiers	Accuracy	Precision	Recall	F1 score
LR_CharLevel_Vectors	0.8031	0.8034	**0.9994**	0.8907
LR_Count_Vectors	**0.8036**	0.8038	**0.9994**	**0.8910**
LR_N-Gram_Vectors	0.8026	0.8030	**0.9994**	0.8905
LR_WordLevel_TF-IDF	0.8031	0.8034	**0.9994**	0.8907
NB_CharLevel_Vectors	0.7886	0.8022	0.9778	0.8813
NB_Count_Vectors	0.8008	0.8052	0.9918	0.8888
NB_N-Gram_Vectors	0.8026	0.8030	**0.9994**	0.8905
NB_WordLevel_TF-IDF	0.8012	**0.8053**	0.9924	0.8891
RF_Count_Vectors	0.7984	0.8050	0.9883	0.8873
RF_WordLevel_TF-IDF	0.7993	0.8052	0.9895	0.8879
SVM_N-Gram_Vectors	0.8026	0.8030	**0.9994**	0.8905
Xgb_CharLevel_Vectors	0.7942	0.8045	0.9825	0.8846
Xgb_Count_Vectors	0.8017	0.8034	0.9971	0.8898
Xgb_WordLevel_TF-IDF	0.8017	0.8034	0.9971	0.8898

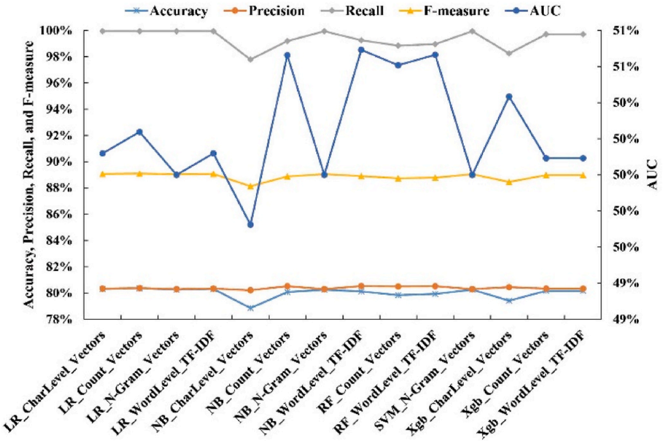


Fig. 4. Evaluation of the 14 classifiers.

Table 1 and Fig. 4 display the performance of the 14 classifiers. The findings revealed that LR_Count_Vectors ranked highest in Recall, while NB_WordLevel_TF-IDF excelled in Precision. LR_Count_Vectors achieved the highest Accuracy score. These results suggest minimal variation in classification performance among the 14 classifiers. LR_Count_Vectors demonstrated the highest values for Accuracy, Recall, and F1-score, whereas NB_WordLevel_TF-IDF exhibited the highest Precision value. Consequently, both LR_Count_Vectors and NB_WordLevel_TF-IDF could serve as suitable final classifiers. With NB_WordLevel_TF-IDF, reviews lacking annotated helpfulness labels are predicted. Among these, 92,966 reviews are classified as helpful by the model, which, when combined with the original 6,813 helpful reviews, contributes to a total of 99,779 helpful reviews consisting of 402,188 review sentences from a total of

Table 2
Number of courses and helpful reviews by subject domains.

Names	Number of courses	Proportion	Number of reviews	%
Art & Design	12	2.87	376	0.37
Humanities	51	12.20	6,564	6.51
Computer Science	45	10.77	1,852	1.84
Engineering	23	5.50	1,501	1.49
Programming	38	9.09	41,344	40.99
Health & Medicine	36	8.61	14,606	14.48
Data Science	35	8.37	810	0.80
Mathematics	16	3.83	1,116	1.11
Science	32	7.66	2,158	2.14
Business	54	12.92	3,476	3.45
Education & Teaching	24	5.74	3,966	3.93
Personal Development	17	4.07	21,083	20.90
Social Sciences	35	8.37	2,008	1.99
Total	418	100	100,860	100

401 courses.

Table 2 provides an overview of the number of courses and helpful reviews categorized by subject. It should be noted that the total numbers of courses and helpful reviews in Table 2 sum to 418 and 100,860, respectively. These figures are higher than 401 and 99,779 due to some courses being classified by the Class Central platform under multiple discipline domains. For example, a course named “Computational Social Science” belongs to both the Data Science and Social Sciences domains. Therefore, these courses are counted in each relevant domain, resulting in some duplication in the tally of courses and helpful reviews.

5.2. Online course review topic identification using BERT models

The experimental dataset comprises 11,732 review sentences used to compare the performance of BERT models and baseline models in review topic classification. The dataset is split into 60% (7,042) for training, 20% (2,359) for validation, and 20% (2,340) for testing, with the allocation done randomly. The standard deviation of 1142.2 suggests variation in sample sizes across categories. For instance, the “Instructor” category has 1,458 records, while the “Relationship” category contains only 191 records. Table 3 presents the quantitative statistics of the experimental dataset. The datasets for training, validation, and testing, which make up 60%, 20%, and 20% of the total dataset, respectively, were selected randomly.

Table 4 compares the performance of the fine-tuned BERT classifier against six baseline models on the same experimental dataset. The results show that the fine-tuned BERT model outperforms the baseline models in predicting topic categories for MOOC learner-generated reviews. Specifically, it achieves a Precision score of 0.784, a Recall score of 0.744, and an F1 score of 0.759. In contrast, the BERT model without fine-tuning demonstrates slightly lower Precision and F1 scores, at 0.776 and 0.757, respectively, though these values still surpass most of the baseline models. Among the Word2Vec-based baselines, the Word2vec + FastText model performs the best in terms of Precision. The Word2vec + HAN model achieves the highest Recall score of 0.749, outperforming all models, including the fine-tuned BERT model. Additionally, the Word2vec + SANN model attains the highest F1 score of 0.758 among the baseline models.

Table 5 presents the performance of the fine-tuned BERT model across various categories. Notably, the categories “Instructor”, “Assessment”, and “Process” achieved Precision scores of 0.8942, 0.8642, and 0.8398, respectively, ranking highest in this metric. In terms of Recall, the top three categories were “Instructor”, “Learning Resources”, and “Others”, with scores of 0.9276, 0.8594, and 0.8311, respectively. Similarly, for F1 scores, “Instructor”, “Assessment”, and “Others” had the highest values of 0.9106, 0.8438, and 0.8342, respectively. Notably, the “Instructor” category demonstrated the highest accuracy score of 92.76% among all categories. Overall, the fine-tuned BERT model showed strong performance in accurately identifying several categories, including “Instructor”, “Learning Resources”, “Others”, and “Assessment”. However, it performed less effectively in categories such as “Learner” and “Relationship”.

We also illustrated the effectiveness of the fine-tuned BERT model across the 10 categories using a confusion matrix (see Fig. 5). Among these categories, “Instructor” exhibited the highest level of agreement between coders and the model, with a consistency percentage of 0.928. Additionally, “Learning Resources”, “Others”, and “Assessment” showed commendable agreement rates of 0.859, 0.831, and 0.824, respectively. However, the categories “Relationship” and “Learner” displayed lower levels of agreement and were often misclassified as “Course Quality” or “Others”.

Table 3
Quantitative statistics of the experimental dataset.

Categories	Training dataset	Validation dataset	Testing dataset	Total
Course introduction	455	142	146	743
Course quality	1,718	609	653	2,980
Learning resources	335	123	127	585
Instructor	901	272	285	1,458
Learner	524	157	126	807
Relationship	125	32	34	191
Assessment	377	146	123	646
Process	187	73	42	302
Platforms and tools	341	107	91	539
Others	2,079	689	713	3,481
Total	7,042	2,350	2,340	11,732
Means	704	235	234	1,173
Standard deviation	669.1	227.6	246.8	1142.2

Table 4
Comparison of the BERT models with baselines.

Models	Precision	Recall	F1 score
Word2vec + CNN	0.761	0.741	0.751
Word2vec + FastText	**0.784**	0.719	0.750
Word2vec + CRNN	0.744	0.747	0.746
Word2vec + HAN	0.761	**0.749**	0.755
Word2vec + RNN	0.765	0.714	0.739
Word2vec + SANN	0.772	0.744	0.758
BERT before finetune	0.776	0.744	0.757
BERT after finetune	**0.784**	0.744	**0.759**

Table 5
Performance of the fine-tuned BERT model across categories.

Categories	Precision	Recall	F1 Score
Course introduction	0.7412	0.6658	0.7015
Course quality	0.7722	0.7993	0.7855
Learning resources	0.6928	0.8594	0.7671
Instructor	**0.8942**	**0.9276**	**0.9106**
Learner	0.6032	0.5584	0.5799
Relationship	0.7229	0.4000	0.5150
Assessment	0.8642	0.8244	0.8438
Process	0.8398	0.6407	0.7269
Platforms and tools	0.7331	0.7291	0.7311
Others	0.8374	0.8311.	0.8342

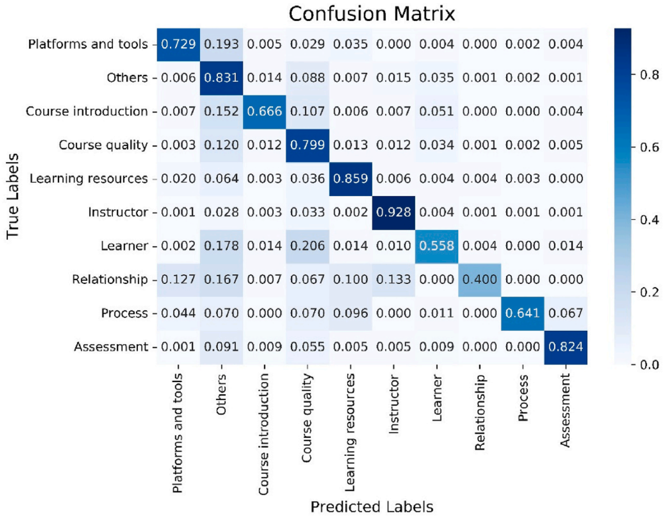


Fig. 5. Confusion matrix of the fine-tuned BERT model.

5.3. *Influencing factors for MOOC learners' satisfaction*

The fine-tuned BERT classifier was used to automatically identify the topic category of the 390,456 unlabeled review sentences, thus generating a finalized dataset with 402,188 review sentences. The distribution of various topics within this dataset is depicted in Fig. 6. Notably, the "Others" category contains the highest number of reviews, while "Assessment" has the fewest. Based on our examination of the dataset, a significant portion of the reviews in the "Others" category consists of general comments on learners' overall perceptions of the MOOC (e.g., "Overall, it was a good course") or expressions of appreciation and recommendation (e.g., "I would recommend it to all ambitious young adults"). These comments provide positive feedback or endorsements but lack specific details about aspects such as course content, teaching methods, or technical elements of the MOOC or the learning processes. Consequently, they were grouped into the "Others" category to maintain the clarity and precision of the classification framework.

Table 6 presents the results of the multiple linear regression analysis across the entire dataset of 402,188 review sentences. The F-test confirms the statistical significance of the model ($F(9, 99767) = 351.9$, $p < 0.001$). The analysis reveals that the frequencies of "Instructor" ($\beta = 0.045$, $p < 0.001$) and "Learner" ($\beta = 0.021$, $p < 0.001$) are positively associated with learner satisfaction. Conversely, the frequencies of "Platforms and Tools" ($\beta = -0.053$, $p < 0.001$), "Course Introduction" ($\beta = -0.041$, $p < 0.001$), "Learning Resources" ($\beta = -0.028$, $p < 0.001$), "Relationship" ($\beta = -0.019$, $p < 0.01$), "Process" ($\beta = -0.119$, $p < 0.001$), and "Assessment" ($\beta = -0.121$, $p < 0.001$) are negatively associated with learner satisfaction.

The comparative analysis of regression results across various subject domains, each containing over 1000 course reviews, is outlined in Tables 7–9 (for Programming, Personal Development, and Health and Medicine courses, respectively) and Tables A2–A9 (for Humanities, Education & Teaching, Business, Science, Social Science, Computer Science, Engineering, and Mathematics courses) in Appendix. Table 7 presents summary statistics regarding the variables in the multiple linear regression model for Programming courses. The F-test confirms the statistical significance of the model ($F(9, 41308) = 158.6$, $p < 0.001$). The findings reveal a positive association between the frequency of "Instructor" ($\beta = 0.051$, $p < 0.001$) and learners' overall satisfaction with programming courses. Conversely, the frequencies of "Platforms and Tools" ($\beta = -0.033$, $p < 0.001$), "Course Introduction" ($\beta = -0.020$, $p < 0.001$), "Course Quality" ($\beta = -0.019$, $p < 0.001$), "Learning Resources" ($\beta = -0.023$, $p < 0.001$), "Relationship" ($\beta = -0.035$, $p < 0.01$), "Process" ($\beta = -0.114$, $p < 0.001$), and "Assessment" ($\beta = -0.096$, $p < 0.001$) are negatively associated with learners' overall satisfaction with programming courses.

Table 6
Multiple linear regression analysis results.

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.841	0.003	1742.931***	0.000
Course introduction	−0.041	0.004	−11.197***	0.000
Course quality	−0.001	0.002	−0.547	0.584
Learning resources	−0.028	0.004	−8.378***	0.000
Instructor	0.045	0.002	19.168***	0.000
Learner	0.021	0.004	5.819***	0.000
Relationship	−0.019	0.007	−2.667**	0.008
Assessment	−0.121	0.003	−37.866***	0.000
Process	−0.119	0.009	−13.452***	0.000
Platforms and tools	−0.053	0.004	−13.014***	0.000

Table 7
Multiple linear regression analysis results (Programming courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.866	0.004	1240.081***	0.000
Course introduction	−0.020	0.005	−3.916***	0.000
Course quality	−0.019	0.002	−8.168***	0.000
Learning resources	−0.023	0.005	−4.41***	0.000
Instructor	0.051	0.003	17.752***	0.000
Learner	−0.003	0.007	−0.388	0.698
Relationship	−0.035	0.013	−2.637**	0.008
Assessment	−0.096	0.004	−23.614***	0.000
Process	−0.114	0.013	−8.952***	0.000
Platforms and tools	−0.033	0.006	−5.605***	0.000

Table 8
Multiple linear regression analysis results (Personal Development courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.886	0.004	1171.005***	0.000
Course introduction	−0.012	0.007	−1.627	0.104
Course quality	0.002	0.003	0.823	0.410
Learning resources	−0.010	0.005	−1.864	0.062
Instructor	0.028	0.004	6.873***	0.000
Learner	0.014	0.005	2.874**	0.004
Relationship	−0.027	0.017	−1.652	0.099
Assessment	−0.074	0.007	−10.99***	0.000
Process	−0.072	0.021	−3.525***	0.000
Platforms and tools	−0.051	0.007	−7.527***	0.000

Table 9
Multiple linear regression analysis results (Health & Medicine courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.832	0.006	746.081***	0.000
Course introduction	0.007	0.008	0.860	0.390
Course quality	0.011	0.004	3.188**	0.001
Learning resources	−0.003	0.008	−0.311	0.756
Instructor	0.026	0.010	2.502*	0.012
Learner	0.023	0.006	4.026***	0.000
Relationship	0.015	0.015	1.050	0.294
Assessment	−0.060	0.012	−5.155***	0.000
Process	−0.039	0.025	−1.551	0.121
Platforms and tools	−0.041	0.009	−4.367***	0.000

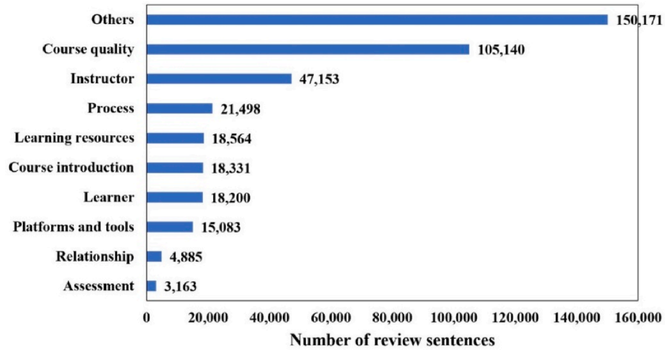


Fig. 6. Number of review sentences in different topics.

Table 8 shows summary statistics for the variables in the multiple linear regression model for Personal Development courses. The F-test confirms the statistical significance of the model ($F(9, 21058) = 33.87$, $p < 0.001$). The analysis reveals a positive association between the frequencies of "Instructor" ($\beta = 0.028$, $p < 0.001$) and "Learner" ($\beta = 0.014$, $p < 0.001$) with the overall satisfaction of learners in Personal Development courses. Conversely, the frequencies of "Platforms and Tools" ($\beta = -0.051$, $p < 0.001$), "Process" ($\beta = -0.072$, $p < 0.001$), and "Assessment" ($\beta = -0.074$, $p < 0.001$) exhibit negative correlations with learners' overall satisfaction with Personal Development courses.

Table 9 shows summary statistics for the variables in the multiple linear regression model for Health and Medicine courses. The F-test confirms the statistical significance of the model ($F(9, 14595) = 9.938$, $p < 0.001$). The analysis reveals a positive relationship between the frequencies of "Course Quality" ($\beta = 0.011$, $p < 0.001$), "Instructor" ($\beta = 0.026$, $p < 0.05$), and "Learner" ($\beta = 0.023$, $p < 0.001$) with learners' overall satisfaction with Health and Medicine courses. However, the frequencies of "Platforms and Tools" ($\beta = -0.041$, $p < 0.001$) and "Assessment" ($\beta = -0.060$, $p < 0.001$) demonstrate negative correlations with learners' overall satisfaction with Health and Medicine courses.

6. Discussion

6.1. Validity of the review helpfulness classification model (RQ1)

Reviews are increasingly recognized as a vital tool on online commerce platforms for influencing potential users' purchasing decisions, and helpful reviews simplify the decision-making process for customers contemplating whether to buy a product or service (Elhadidy, 2017; Von Helversen et al., 2018). We addressed the challenge of evaluating extensive course review data by employing and validating machine learning classifiers. Results from the performance evaluation of these classifiers (see Table 1) indicated their effectiveness in automatically classifying the helpfulness of MOOC reviews. The performance of all 14 classifiers appeared similar, particularly concerning metrics such as Accuracy, Precision, Recall, and F-score. Overall, the machine learning models performed on par with human coders, underscoring their potential in assessing review helpfulness.

From a methodological standpoint, this work enhanced prior studies (e.g., Lubis et al., 2017, 2019) by incorporating a wider array of machine learning algorithms and covering MOOCs from diverse subject areas. Consequently, the model developed in this study demonstrated greater generalizability. The process of utilizing machine learning to identify helpful reviews laid the groundwork for generating input data for advanced analyses concerning topic classification and the impact of review topics on learner satisfaction.

The process of identifying review helpfulness is crucial for the intelligence-driven operations of MOOCs, as reviews contain valuable insights into learners' opinions. The absence or misapplication of review helpfulness prediction methods can lead to ambiguous findings (Chen et al., 2022). Additionally, this approach offers significant advantages to MOOC providers and managers by supporting the design, development, and implementation of marketing and operational strategies (Chen et al., 2024). Integrating this study's approach into their platforms can help establish guidelines for crafting informative and impactful reviews. By using an effective model for identifying helpful reviews, MOOC providers can enhance their ability to develop comprehensive MOOC intelligence systems, thereby gaining deeper insights into learners' preferences and the key attributes highlighted in reviews (Lee & Choeh, 2014). Ultimately, this enables MOOC designers to leverage these systems to improve the efficiency of course implementation and management efforts (Mubarak et al., 2021).

6.2. Validity of the review topic classification model (RQ2)

This study employed a fine-tuned BERT classifier for review topic classification, marking a significant advancement compared to conventional machine learning methods that relied solely on separate semantic elements, such as the bag-of-words model and shallow neural networks. Prior research often utilized text mining or linguistic features in conjunction with machine learning algorithms to delineate learners' review topic categories (Okoye et al., 2022). However, these methods exhibit certain limitations when compared to the fine-tuned BERT model. This disparity may stem from the fact that conventional machine learning models require feature engineering to extract crucial features or reduce dimensionality, which can result in the loss of predictive information (Zheng et al., 2014).

Through the comparison of fine-tuned BERT with other models (see Table 4), this study determined that fine-tuned BERT outperformed the other models in review topic classification. Several factors contribute to the efficacy of fine-tuned BERT. First, BERT models undergo pre-training on extensive and diverse datasets in an unsupervised manner (Wang et al., 2023). During this initial phase, the model assimilates intricate, contextualized representations of words, capturing nuanced language patterns and relationships. Second, fine-tuning adapts the pre-trained BERT model to the specific features and nuances of the downstream task—in this case, course review topic classification. This process enables the model to apply its learned knowledge from the pre-training phase to the task at hand, effectively transferring its understanding of language intricacies and structures (Zhao et al., 2023). The fine-tuned BERT model effectively addressed challenges such as sparse and unbalanced manual coding datasets. In human coding, the time spent on coding tasks may vary depending on the coder's familiarity with the content and categories. Typically, human coding takes around five to 10 s per sentence, based on sample data coding experiences. In contrast, automated coding performed by machines is more cost-effective and scalable (Chen et al., 2024) in handling large volumes of data. The findings underscore the potential of the fine-tuned BERT model in educational contexts.

6.3. Influential factors for MOOC learner satisfaction (RQ3)

The multiple linear regression analysis provides insights into the various factors influencing MOOC learners' satisfaction across different subject domains. Regarding the comprehensive evaluation of MOOC satisfaction, the collective analysis findings (see Table 6) reveal that the frequency of "Instructor" is positively associated with overall satisfaction, underscoring the universal significance of effective pedagogy (Schallert et al., 2015). Learners across diverse fields highly value instructors who are knowledgeable, approachable, and engaging (Hew et al., 2020; Watson et al., 2017). Additionally, our analysis indicates that learners expressing higher levels of overall satisfaction tend to provide more personal information (e.g., background, interests, needs, and experiences) when giving feedback on the MOOCs they have enrolled in. This aligns with Bayeck's (2016) and Milligan and Littlejohn's (2017) findings that motivations for enrolling in MOOC courses include supplementing formal education, preparing for future endeavors, and enhancing knowledge and skills. However, there are notable variations across different study areas. For instance, most students enrolled in humanities-related MOOCs out of curiosity, while most students in social sciences-related MOOCs did so to enhance work performance. Therefore, MOOC providers are advised to customize students' experiences by tailoring content and pathways according to their motivations and requirements.

Conversely, negative associations were observed with the frequencies of.

"Platforms and tools", "Course introduction", "Learning resources", "Relationships", "Process", and "Assessment", suggesting common areas of concern. Issues related to platform functionality, course content introduction, and the learning process universally detract from satisfaction (e.g., Deng et al., 2019; Deshpande & Chukhlomin, 2017; Gupta, 2021). These findings suggest that learners' dissatisfaction primarily stems from the subpar performance or quality of these aspects. Upon examining the regression outcomes across various subject domains (see Tables 7–9 and Tables A2–A9 in the **Appendix**), it is evident that the frequency of "Assessment" significantly affects overall satisfaction in 9 out of 11 subject domains. Assessment, which has been widely reported to positively influence MOOC learners' satisfaction (e.g., Hew et al., 2020; Jordan, 2015), should go beyond mere knowledge recall and provide learners with opportunities for application (Bali, 2014). Furthermore, the frequency of "Instructor" is found to have a positive influence on overall satisfaction in 8 out of 11 subject domains, suggesting that satisfied learners frequently discuss aspects related to instructors (Hew, 2018).

Moreover, the frequency of "Platforms and tools" was identified as negatively impacting overall satisfaction in 7 out of 11 subject domains. Representative reviews from learners are presented in Table A10 in the **Appendix**. Dissatisfied students often cited issues related to the substandard quality of videos, forums, and auto-graders, which echoes Chen et al. (2024)'s findings. These tools play a critical role in the MOOC experience, and any shortcomings can significantly impact learner satisfaction. First, learners highlighted several issues with the videos, including the lack of animations to demonstrate concepts such as stepping through code line-by-line and setting breakpoints. Common concerns also included the absence of close-up shots during interviews and poor coherence in video transcripts. These issues likely stem from inadequate production quality and a lack of attention to detail in content delivery (Watson et al., 2017). As videos are a primary medium for delivering content in MOOCs, such deficiencies can lead to disengagement and frustration, reducing the effectiveness of the learning experience, as supported by Shukor and Abdullah (2019).

Second, discussion forums, another critical component of MOOCs, also faced considerable criticism (Li et al., 2018). Learners reported that forums were not user-friendly, were inundated with low-quality posts, and lacked a sense of community. Many users expressed frustration about the absence of meaningful interaction with instructors or peers, as well as the poor design and navigation features of the forums. These issues undermine the forums' intended role in fostering instructional support and peer engagement. A poorly designed or moderated forum can leave learners feeling isolated and unsupported, negatively affecting their overall satisfaction, retention, and completion rates (e.g., Coetzee et al., 2014; Lee et al., 2011). In addition, auto-graders, particularly important in programming-related MOOCs (Staubitz et al., 2015), also received significant criticism. Many learners reported strict output-matching requirements, buggy behavior, and incorrect evaluations of correct answers. Such issues frustrate learners, especially when trivial deviations or technical glitches prevent them from receiving credit for valid work. As auto-graders are integral to scalable hands-on programming assignments, their flaws can severely impact learner satisfaction and trust in the platform (Chen et al., 2021).

To address these issues, several improvements can be implemented. For videos, animations and interactive content should be introduced to explain complex concepts more effectively. Enhancing production quality through better editing, close-up shots, and refined transcripts would significantly improve the learning experience. For forums, redesigning their layout to improve navigation and search functionality, coupled with active moderation to filter low-quality posts, can foster a stronger sense of community. Encouraging instructor participation and peer mentoring programs would further enhance engagement and create a more supportive environment. To improve auto-graders, their operations should be visualized to give learners a clearer understanding of how their submissions are evaluated. Allowing inline discussions for feedback on auto-grader evaluations can help resolve issues collaboratively. Additionally, increasing the tolerance for minor deviations in output and regularly debugging the system to ensure accuracy and fairness would address many of the learners' concerns. These targeted improvements would not only enhance learner satisfaction but also improve the overall effectiveness and reliability of MOOCs.

Furthermore, the frequency of "Process" was observed to negatively influence overall satisfaction in 5 out of 11 subject domains. Representative reviews from learners, provided in Table A11 in the **Appendix**, reveal common themes of dissatisfaction related to problem-solving, feedback, and the lack of real-world examples or case studies, which aligns with Shah et al. (2022)'s findings. These issues hindered learners' ability to effectively engage with course materials and impacted their overall satisfaction with the learning process (Hew, 2018). Specifically, one recurring issue highlighted by learners is the inadequacy of feedback provided during the course (Warren et al., 2014, pp. 665–670). Many reviews noted that responses from senators often appeared to be generic and resembled textbook-like explanations that failed to address practical problems. This lack of personalized, contextualized support left learners feeling unsupported and frustrated. The absence of consistent and timely feedback from instructors or mentors compounded the issue, leaving learners uncertain about their progress and unsure whether they were heading in the right direction (Hew & Cheung, 2014). Such gaps in feedback mechanisms undermined their confidence and satisfaction.

Another significant factor contributing to dissatisfaction was the insufficient depth and comprehensiveness of the lecture content, which has also been reported by Aparicio et al. (2019) and Hone and El Said (2016). Several learners reported that they had to rely heavily on external resources, such as Stack Overflow, to solve problems because the course materials did not provide enough detail or clarity. This reliance on external help added to the cognitive and emotional burden of the learning process and detracted from the course experience (Peng & Xu, 2020). Learners expressed frustration with the expectation to solve problems without being equipped with the necessary information through lectures, which often resulted in wasted effort and confusion (Qaddumi et al., 2021). Additionally, the abstract nature of the principles taught further exacerbated learner dissatisfaction. While foundational principles were often explained in simple terms, their application in complex, real-life contexts was rarely addressed (Conrad & Openo, 2018). Reviews indicated that the examples or analogies provided in the course did not always align well with the introduced concepts, creating gaps in understanding. The absence of real-world examples or demonstrations limited learners' ability to grasp the practical relevance of the material and undermined their ability to apply their knowledge effectively (Hew et al., 2020).

To address these issues, several improvements can be made. First, enhancing feedback mechanisms is crucial. Instructors and course assistants should provide timely, detailed, and contextualized feedback that addresses specific challenges faced by learners. Automated systems could also be implemented to offer instant feedback on common issues, while senators should be encouraged to provide clear and structured responses that outline problem-solving steps. Including regular progress reviews or checkpoints could also help learners assess their understanding and receive guidance to stay on track. Second, improving the depth and comprehensiveness of lecture content is necessary. Lectures should be designed to cover topics in sufficient detail to enable learners to solve course problems independently. Interactive problem-solving sessions, where instructors demonstrate step-by-step approaches to tackling challenges, would further enhance comprehension and engagement. Lastly, incorporating real-world examples is essential for bridging the gap between theory and practice. Using relatable case studies, contextual applications, and live demonstrations of principles in action would help learners understand how to apply concepts in practical settings. Supplementing this with an expanded library of examples and explanatory slides would provide learners with additional resources to support their learning journey. By addressing these concerns, MOOC instructors and designers can create a more supportive and engaging learning process. Strengthening feedback mechanisms, providing more comprehensive instructional content, and integrating real-world applications will help learners overcome challenges, manage frustrations, and achieve higher satisfaction with the course.

6.4. Theoretical and practical implications

This work offers both theoretical and practical implications. As the educational landscape undergoes rapid changes, there is a growing demand for precise, data-driven methodologies to understand user needs. Theoretical contributions arise from our nuanced exploration of the multifaceted dynamics that govern online education, facilitated by the sophisticated analytical capabilities of machine learning algorithms and BERT models. Through our in-depth investigation of influential factors and comprehensive analyses, this study not only enriches our understanding of these complex dynamics but also provides deeper insights into how these technologies can uncover underlying patterns and trends in learner interactions within MOOCs.

On a practical level, our findings offer actionable methodologies for educators, course designers, and platform developers. The development of optimal models for automatically classifying review helpfulness and topic categories provides valuable tools for extracting meaningful insights from vast amounts of review data. Moreover, the study delves into the multifaceted factors influencing MOOC learners' overall satisfaction, both across the educational spectrum and within specific subject domains. These anticipated outcomes serve as valuable information for educators, course designers, and platform developers, supporting continuous improvement and optimization of MOOC offerings to better tailor educational experiences to the evolving needs of learners in the digital educational landscape.

In sum, this study demonstrates the transformative potential of machine learning and BERT models in enhancing the quality and accessibility of online education platforms and offers actionable recommendations for stakeholders to improve the quality and effectiveness of MOOCs.

7. Conclusion, contribution, implications, limitations, and future work

This research explores the evolving landscape of MOOCs by investigating the efficacy of machine learning models, particularly BERT, in understanding online course reviews and identifying factors influencing MOOC learners' overall satisfaction. The study makes several significant contributions to the field. First, it demonstrates the effectiveness of machine learning classifiers, including fine-tuned BERT models, in automatically identifying helpful online course reviews. These models achieved performance levels comparable to human coders, underscoring their potential for scalable and efficient assessment of MOOC review helpfulness. Second, this research validates the reliability of fine-tuned BERT in categorizing review topics, offering a nuanced understanding of learner engagement through detailed textual analysis. Compared to traditional machine learning approaches, BERT models showed superior performance, particularly after fine-tuning. Third, insights from multiple linear regression analysis revealed both universal patterns and domain-specific nuances in learner satisfaction. Positive associations between instructor-related factors and satisfaction highlighted the universal importance of effective teaching, while negative associations with platform quality, course content, and learning processes pointed to areas for improvement.

The implications of this study are manifold. Firstly, the developed machine learning models provide MOOC providers with a scalable method for evaluating review usefulness, offering actionable insights for improving course design and implementation. Incorporating these models into platforms can help learners make more informed decisions based on helpful reviews. Secondly, the use of fine-tuned BERT models for topic classification enables MOOC designers to better understand learner preferences and engagement, fostering the creation of tailored learning experiences that address specific learner needs. Lastly, the identification of factors influencing satisfaction informs educators and platform developers about critical areas for improvement, driving the continuous enhancement of MOOC offerings.

Despite its contributions, this study has several limitations. The analysis was based on review data from Class Central, a prominent MOOC aggregator. While this provided a rich dataset, future research could expand the scope by incorporating data from other MOOC platforms such as Coursera, edX, Khan Academy, and FutureLearn, as well as regional platforms like iCourse in Mainland China. This would enable a more comprehensive understanding of learner satisfaction across diverse platforms and regions. Additionally, the baselines used for machine learning comparison may not reflect the latest advancements in large language models (LLMs). While the primary aim of this research was not to benchmark cutting-edge methodologies, future work could explore the integration of state-of-the-art techniques, including LLMs, to further enhance classification performance. Another limitation lies in the dataset composition, particularly the large number of reviews categorized as "Others". These reviews often consisted of general comments on overall MOOC experiences or expressions of appreciation, which lacked specificity in addressing actionable aspects of the courses. While grouping them into a general category maintained clarity in the classification framework, it limited the granularity of insights. Future research could develop additional sub-categories to capture nuanced themes within these general comments, providing more detailed feedback for course improvement.

To conclude, this study addresses critical gaps in MOOC research by integrating machine learning and BERT models for analyzing learner reviews and identifying influential satisfaction factors. It underscores the importance of leveraging advanced analytical tools to enhance the quality and effectiveness of online education. Future research should continue to refine these approaches, broaden data sources, and explore emerging technologies to better serve the evolving needs of learners in the digital education era.

CRediT authorship contribution statement

Xieling Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Haoran Xie:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Di Zou:** Writing – review & editing, Supervision, Project administration, Methodology, Data curation. **Gary Cheng:** Writing – review & editing, Visualization, Supervision, Methodology. **Xiaohui Tao:** Writing – review & editing, Validation, Supervision. **Fu Lee Wang:** Writing – review & editing, Supervision, Methodology, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *Computers & Education: Artificial Intelligence* and was not involved in the editorial review or the decision to publish this article.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62307010) and the Philosophy and Social Science Planning Project of Guangdong Province of China (No. GD24XJY17), as well as Lam Woo Research Fund (LWP20019) and the Faculty Research Grants (DB23B2 and DB24A4) of Lingnan University, Hong Kong. This article is partly based on the first author Xieling Chen's unpublished PhD dissertation entitled "Exploring factors affecting learner satisfaction toward MOOCs: A hybrid approach of topic modeling, deep learning, and structural equation modeling" submitted to The Education University of Hong Kong in 2022.

Appendix

Table A1
Coding scheme.

Categories	Descriptions	Examples
Course introduction	Course information, e.g., syllabus, overview, schedule/calendar, requirement, certificate, credential, payment, language	"This section of the course gave a really solid foundation into the introduction of what computational social science is and why it is so effective"
Course quality	Content quality, information quality, course difficulty, knowledge enhancement, beginner friendliness, practicality, usefulness, helpfulness	"The whole course was very informative and I felt that I gained a fair amount of knowledge"
Learning resources	Availability of learning materials, textbooks, notes, handouts; slides, and additional links	"Reference support and reading material are very satisfying"
Instructor	Instructor knowledge, accessibility, enthusiasm for teaching, humor, presentation, instructing pace	"Charles severance is passionate about python and enthusiastic about sharing his knowledge"
Learner	Learner background, leaner interest, educational needs (e.g., job or academic needs)	"I'm a 32-year-old Italian vendor, interested in all faces of the sustainability subject"
Relationship	Peer interaction, leaner-instructor interaction	"It was a good idea to allow users to interact, I like to read comments made by other students"
Assessment	Quizzes, assignments, projects, exercises, tests, experiments, lab activities, grading	"Quizzes are the best part of the course, according to me"
Process	Giving and receiving feedback, participating in learning activities, problem-solving, availability of cases and examples during learning	"It provided a variety of examples and made me experiment a lot"
Platforms and tools	Platform use and system quality; video quality (captions, transcripts, speed, image, sound)	"The quality of the video is awesome in both the type of course"
Others	Learner perception, overall evaluation, appreciation, or recommendation	"I would recommend it to all ambitious young adults"

Table A2
Multiple linear regression analysis results (Humanities courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.740	0.012	402.518***	0.000
Platforms and tools	−0.052	0.015	−3.523***	0.000
Course introduction	0.003	0.014	0.195	0.845
Course quality	0.013	0.008	1.639	0.101
Learning resources	−0.008	0.013	−0.564	0.573
Instructor	0.052	0.012	4.45***	0.000
Learner	0.007	0.014	0.490	0.624
Relationship	0.039	0.018	2.164*	0.030
Process	−0.090	0.032	−2.834**	0.005
Assessment	−0.122	0.017	−7.364***	0.000

Table A3
Multiple linear regression analysis results (Education & Teaching courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.760	0.014	333.425***	0.000
Platforms and tools	−0.037	0.020	−1.842	0.066
Course introduction	−0.011	0.016	−0.683	0.495
Course quality	0.013	0.009	1.387	0.166
Learning resources	0.010	0.016	0.608	0.543
Instructor	0.033	0.011	2.878**	0.004
Learner	0.017	0.017	1.000	0.317
Relationship	−0.020	0.019	−1.049	0.294
Process	−0.042	0.028	−1.492	0.136
Assessment	0.016	0.012	1.353	0.176

Table A4
Multiple linear regression analysis results (Business courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.690	0.018	259.527***	0.000
Platforms and tools	−0.170	0.024	−7.132***	0.000
Course introduction	−0.048	0.021	−2.32*	0.020
Course quality	0.026	0.009	2.801**	0.005
Learning resources	−0.110	0.024	−4.512***	0.000
Instructor	0.092	0.018	5.123***	0.000
Learner	0.092	0.024	3.833***	0.000

(*continued on next page*)

Table A4 (*continued*)

Variables	Coefficients	Std. Error	t value	Sig.
Relationship	0.019	0.046	0.406	0.684
Process	−0.074	0.039	−1.888	0.059
Assessment	−0.117	0.017	−6.803***	0.000

Table A5
Multiple linear regression analysis results (Science courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.778	0.021	226.167***	0.000
Platforms and tools	−0.073	0.028	−2.596**	0.009
Course introduction	−0.032	0.026	−1.234	0.218
Course quality	−0.006	0.013	−0.479	0.632
Learning resources	0.019	0.025	0.764	0.445
Instructor	0.062	0.021	2.934**	0.003
Learner	0.045	0.028	1.647	0.100
Relationship	0.005	0.049	0.107	0.914
Process	0.050	0.058	0.855	0.393
Assessment	−0.180	0.027	−6.673***	0.000

Table A6
Multiple linear regression analysis results (Social Science courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.821	0.027	175.393***	0.000
Platforms and tools	−0.030	0.039	−0.779	0.436
Course introduction	−0.008	0.034	−0.239	0.812
Course quality	−0.018	0.018	−1.024	0.306
Learning resources	−0.035	0.031	−1.143	0.253
Instructor	−0.018	0.030	−0.616	0.538
Learner	0.034	0.038	0.888	0.374
Relationship	−0.023	0.055	−0.416	0.677
Process	−0.106	0.087	−1.215	0.224
Assessment	−0.170	0.037	−4.568***	0.000

Table A7
Multiple linear regression analysis results (Computer Science courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.660	0.034	136.022***	0.000
Platforms and tools	−0.127	0.053	−2.407*	0.016
Course introduction	−0.060	0.038	−1.603	0.109
Course quality	0.012	0.019	0.633	0.527
Learning resources	−0.100	0.039	−2.531*	0.011
Instructor	−0.056	0.036	−1.529	0.126
Learner	0.050	0.060	0.838	0.402
Relationship	0.110	0.094	1.171	0.242
Process	−0.249	0.080	−3.107**	0.002
Assessment	−0.073	0.030	−2.433*	0.015

Table A8
Multiple linear regression analysis results (Engineering courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.760	0.023	208.339***	0.000
Platforms and tools	0.013	0.033	0.401	0.688
Course introduction	−0.010	0.033	−0.294	0.769
Course quality	−0.008	0.012	−0.681	0.496
Learning resources	0.026	0.031	0.854	0.393
Instructor	0.077	0.021	3.671***	0.000
Learner	0.032	0.037	0.859	0.390
Relationship	0.050	0.065	0.762	0.446
Process	−0.050	0.052	−0.954	0.340
Assessment	−0.099	0.030	−3.285**	0.001

Table A9
Multiple linear regression analysis results (Mathematics courses).

Variables	Coefficients	Std. Error	t value	Sig.
Intercept	4.828	0.049	99.128***	0
Platforms and tools	−0.092	0.071	−1.295	0.196
Course introduction	−0.085	0.066	−1.284	0.200
Course quality	0.054	0.030	1.782	0.075
Learning resources	−0.157	0.051	−3.058**	0.002
Instructor	−0.087	0.036	−2.419*	0.016
Learner	−0.139	0.094	−1.487	0.138
Relationship	0.130	0.117	1.114	0.266
Process	−0.417	0.134	−3.124**	0.002
Assessment	−0.029	0.050	−0.592	0.554

Table A10
Representative complaining course reviews regarding “platforms and tools”

No.	Review comments
1	I think the video should show the actual animation of stepping through code line-by-line and setting breakpoints in example software using such a tool
2	The video would, however, be greatly enhanced by including more close-up shots of the interview.
3	If you read the transcript of the video, the sentences spoken would not even make much sense.
4	There were also a few downsides, but, interestingly enough, most of them were technical: problems with the final exam, trading issues, and resolving this mess sure took some time. Other issues included varying sound levels during parts of the video and a total, horrible mess on the discussion forum.
5	The forum is not user-friendly either.
6	I personally don’t like forum boards—I feel lost on them, and they seem almost non-existent.
7	The forum was inundated with low-quality posts, so there was never really a sense of community.
8	There won’t be answers in the video or on the forum.
9	The autograder didn’t allow me to import the patron’s regret library, and I wasn’t given credit for solving the problem.
10	My only complaint was the tools used for the session. The autograder is quite annoying because it requests the exact output as prompted on the screen.
11	The only problem I had was the autograder, as trivial issues in my code prevented me from submitting the material, even though the code was working and displaying the required output.
12	The autograder is sometimes buggy, and code needs to be adjusted to clear false error messages.
13	Sometimes correct answers were incorrectly marked.

Table A11
Representative complaining course reviews regarding “process”

No.	Review comments
1	When “senators” reply to some questions, their answers often seem to be taken straight from a textbook
2	Some senators attempted to address learners’ problems in the course discussion forum. However, these discussions were very confusing and did not clearly identify how to solve the problem
3	The course itself is not terrible, but be prepared to do a lot of searching for outside help on Stack Overflow and similar platforms, as the lectures do not provide sufficient material to solve the problems.
4	They expect you to solve some problems without giving you the necessary information in the lectures.
5	I apologize for the negative feedback, but this is honestly how I felt about the course.
6	I got lost, and since there was no feedback, I often did not know whether I was heading in the right direction.
7	It would have been extremely helpful if someone had provided feedback because there was one hurdle I couldn’t get past, despite spending a lot of time researching the topic.
8	The course also didn’t provide sufficient feedback to help me solve a problem, which was probably related to text encoding, but it was very frustrating nonetheless.
9	There is no feedback from the instructor or senators, who usually respond to posts from students requesting help or information.
10	Principles were explained in a very simple way but were never presented in their more complex, real-life contexts during the class.
11	The examples or analogies provided did not always align with the concepts being introduced.
12	I would have liked to see more examples and further explanations of the components.
13	It is suggested to kindly provide more live examples, which could ease the understanding process, along with slides wherever possible.

References

Albelbisi, N. A. (2020). Development and validation of the MOOC success scale (MOOC-SS). *Education and Information Technologies, 25*(5), 4535–4555. https://doi.org/10.1007/s10639-020-10186-4

Alraimi, K. M., Zo, H., & Ciganek, A. P. (2015). Understanding the MOOCs continuance: The role of openness and reputation. *Computers & Education, 80*, 28–38. https://doi.org/10.1016/j.compedu.2014.08.006

Alsayat, A., & Ahmadi, H. (2023). A hybrid method using ensembles of neural network and text mining for learner satisfaction analysis from big datasets in online learning platform. *Neural Processing Letters, 55*(3), 3267–3303. https://doi.org/10.1007/s11063-022-11009-y

Aparicio, M., Oliveira, T., Bacao, F., & Painho, M. (2019). Gamification: A key determinant of massive open online course (MOOC) success. *Information & Management, 56*(1), 39–54. https://doi.org/10.1016/j.im.2018.06.003

Bali, M. (2014). MOOC pedagogy: Gleaning good practice from existing MOOCs. *Journal of Online Learning and Teaching, 10*(1), 44. Retrieved January 5, 2024 from https://jolt.merlot.org/vol10no1/bali_0314.pdf?ref=hybrid-pedagogy.

Bayeck, R. (2016). Exploratory study of MOOC learners’ demographics and motivation: The case of students involved in groups. *Open Praxis, 8*(3), 223–233. International Council for Open and Distance Education. Retrieved January 5, 2024 from https://www.learntechlib.org/p/173534/.

Bitakou, E., Ntaliani, M., Demestichas, K., & Costopoulou, C. (2023). Assessing massive open online courses for developing digital competences among higher education teachers. *Education Sciences, 13*(9), 900. https://doi.org/10.3390/educsci13090900

Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., & Gašević, D. (2020). How good is my feedback? A content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 428–437). https://doi.org/10.1145/3375462.3375477

Chakraborty, U., & Biswal, S. K. (2023). Are online reviews credible? The effects of online reviews for the adoption of MOOCs for E-learning. *Journal of Decision Systems, 32*(4), 678–699. https://doi.org/10.1080/12460125.2022.2133370

Chen, X., Cheng, G., Xie, H., Chen, G., & Zou, D. (2021). Understanding MOOC reviews: Text mining using structural topic model. *Human-Centric Intelligent Systems, 1*(3–4), 55–65. https://doi.org/10.2991/hcis.k.211118.001

Chen, X., Wang, F. L., Cheng, G., Chow, M.-K., & Xie, H. (2022). Understanding learners' perception of MOOCs based on review data analysis using deep learning and sentiment analysis. *Future Internet, 14*(8), 218. https://doi.org/10.3390/fi14080218

Chen, X., Zou, D., Cheng, G., & Xie, H. (2024). Deep neural networks for the automatic understanding of the semantic content of online course reviews. *Education and Information Technologies, 29*(4), 3953–3991. https://doi.org/10.1007/s10639-023-11980-6

Chen, X., Zou, D., Xie, H., & Cheng, G. (2020). What are MOOCs learners' concerns? Text analysis of reviews for computer science courses. In *International conference on database systems for advanced applications* (pp. 73–79). Springer International Publishing. https://doi.org/10.1007/978-3-030-59413-8_6.

Coetzee, D., Fox, A., Hearst, M. A., & Hartmann, B. (2014). Should your MOOC forum use a reputation system?. In *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing* (pp. 1176–1187). https://doi.org/10.1145/2531602.2531657

Conrad, D., & Openo, J. (2018). *Assessment strategies for online learning: Engagement and authenticity*. Athabasca University Press.

Darmawansah, D., Hwang, G.-J., Chen, M.-R. A., & Liang, J.-C. (2023). Trends and research foci of robotics-based STEM education: A systematic review from diverse angles based on the technology-based learning model. *International Journal of STEM Education, 10*(1), 12. https://doi.org/10.1186/s40594-023-00400-3

Deng, R., Benckendorff, P., & Gannaway, D. (2019). Progress and new directions for teaching and learning in MOOCs. *Computers & Education, 129*, 48–60. https://doi.org/10.1016/j.compedu.2018.10.019

Deshpande, A., & Chukhlomin, V. (2017). What makes a good MOOC: A field study of factors impacting student motivation to learn. *American Journal of Distance Education, 31*(4), 275–293. https://doi.org/10.1080/08923647.2017.1377513

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

Elhadidy, D. Y. (2017). To investigate how e-WOM affects young buyers purchasing decision in FMCGS. *Business and Management Review, 8*(5), 252–257. Retrieved January 5, 2024 from https://cberuk.com/cdn/conference_proceedings/conference_41741.pdf.

Fang, J.-W., Hwang, G.-J., & Chang, C.-Y. (2022). Advancement and the foci of investigation of MOOCs and open online courses for language learning: A review of journal publications from 2009 to 2018. *Interactive Learning Environments, 30*(7), 1351–1369. https://doi.org/10.1080/10494820.2019.1703011

Ganguly, B., Sengupta, P., & Biswas, B. (2024). What are the significant determinants of helpfulness of online review? An exploration across product-types. *Journal of Retailing and Consumer Services, 78*, Article 103748. https://doi.org/10.1016/j.jretconser.2024.103748

Gupta, K. P. (2021). Understanding learners' completion intention of massive open online courses (MOOCs): Role of personality traits and personal innovativeness. *International Journal of Educational Management, 35*(4), 848–865. https://doi.org/10.1108/IJEM-01-2020-0042

Gupta, S., Kumar, P., & Tekchandani, R. K. (2023). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimedia Tools and Applications, 82*(8), 11365–11394. https://doi.org/10.1007/s11042-022-13558-9

Hew, K. F. (2018). Unpacking the strategies of ten highly rated MOOCs: Implications for engaging students in large online courses. *Teachers College Record, 120*(1), 1–40. https://doi.org/10.1177/016146811812000107

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review, 12*, 45–58. https://doi.org/10.1016/j.edurev.2014.05.001

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education, 145*, Article 103724. https://doi.org/10.1016/j.compedu.2019.103724

Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers & Education, 98*, 157–168. https://doi.org/10.1016/j.compedu.2016.03.016

Hwang, G., & Fu, Q. (2020). Advancement and research trends of smart learning environments in the mobile era. *International Journal of Mobile Learning and Organisation, 14*(1), 114–129. https://doi.org/10.1504/IJMLO.2020.103911

Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. *International Review of Research in Open and Distributed Learning, 16*(3), 341–358. https://doi.org/10.19173/irrodl.v16i3.2112

Kaur, K., & Kaur, P. (2023). Improving BERT model for requirements classification by bidirectional LSTM-CNN deep model. *Computers & Electrical Engineering, 108*, Article 108699. https://doi.org/10.1016/j.compeleceng.2023.108699

Kong, J., & Lou, C. (2023). Do cultural orientations moderate the effect of online review features on review helpfulness? A case study of online movie reviews. *Journal of Retailing and Consumer Services, 73*, Article 103374. https://doi.org/10.1016/j.jretconser.2023.103374

Koufakou, A. (2024). Deep learning for opinion mining and topic classification of course reviews. *Education and Information Technologies, 29*(3), 2973–2997. https://doi.org/10.1007/s10639-023-11736-2

Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications, 41*(6), 3041–3046. https://doi.org/10.1016/j.eswa.2013.10.034

Lee, P.-J., Hu, Y.-H., & Lu, K.-T. (2018). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics, 35*(2), 436–445. https://doi.org/10.1016/j.tele.2018.01.001

Lee, S. J., Srinivasan, S., Trail, T., Lewis, D., & Lopez, S. (2011). Examining the relationship among student perception of support, course satisfaction, and learning outcomes in online learning. *The Internet and Higher Education, 14*(3), 158–163. https://doi.org/10.1016/j.iheduc.2011.04.001

Li, J., Tang, Y., Cao, M., & Hu, X. (2018). The moderating effects of discipline on the relationship between asynchronous discussion and satisfaction with MOOCs. *Journal of Computers in Education, 5*(3), 279–296. https://doi.org/10.1007/s40692-018-0112-2

Li, Y., Xu, Z., Wang, X., & Wang, X. (2020). A bibliometric analysis on deep learning during 2007–2019. *International Journal of Machine Learning and Cybernetics, 11*, 2807–2826. https://doi.org/10.1007/s13042-020-01152-0

Liu, H., Chen, X., & Zhao, F. (2024). Learning behavior feature fused deep learning network model for MOOC dropout prediction. *Education and Information Technologies, 29*(3), 3257–3278. https://doi.org/10.1007/s10639-023-11960-w

Liu, Z., Kong, X., Chen, H., Liu, S., & Yang, Z. (2023). MOOC-BERT: Automatically identifying learner cognitive presence from MOOC discussion data. *IEEE Transactions on Learning Technologies, 16*(4), 528–542. https://doi.org/10.1109/TLT.2023.3240715

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management, 47*, 140–151. https://doi.org/10.1016/j.tourman.2014.09.020

Liu, S., Peng, X., Cheng, H. N. H., Liu, Z., Sun, J., & Yang, C. (2019). Unfolding sentimental and behavioral tendencies of learners' concerned topics from course reviews in a MOOC. *Journal of Educational Computing Research, 57*(3), 670–696. https://doi.org/10.1177/0735633118757181

Lubis, F. F., Rosmansyah, Y., & Supangkat, S. H. (2017). Improving course review helpfulness prediction through sentiment analysis. In *2017 international conference on ICT for smart society (ICISS)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICTSS.2017.8288877.

Lubis, F. F., Rosmansyah, Y., & Supangkat, S. H. (2019). Topic discovery of online course reviews using LDA with leveraging reviews helpfulness. *International Journal of Electrical and Computer Engineering, 9*(1), 426. https://doi.org/10.11591/ijece.v9i1.pp426-438

Meek, S., Wilk, V., & Lambert, C. (2021). A big data exploration of the informational and normative influences on the helpfulness of online restaurant reviews. *Journal of Business Research, 125*, 354–367. https://doi.org/10.1016/j.jbusres.2020.12.001

Milligan, C., & Littlejohn, A. (2017). Why study on a MOOC? The motives of students and professionals. *International Review of Research in Open and Distributed Learning, 18*(2), 92–102. https://doi.org/10.19173/irrodl.v18i2.3033

Mubarak, A. A., Cao, H., & Ahmed, S. A. M. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. *Education and Information Technologies, 26*(1), 371–392. https://doi.org/10.1007/s10639-020-10273-6

Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Achem, J. A. G., Escamilla, J., & Hosseini, S. (2022). Towards teaching analytics: A contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification. *Education and Information Technologies, 27*, 3891–3933. https://doi.org/10.1016/10.1007/S10639-021-10751-5

O'Mahony, M. P., & Smyth, B. (2009). Learning to recommend helpful hotel reviews. In *Proceedings of the third ACM conference on recommender systems* (pp. 305–308). https://doi.org/10.1145/1639714.1639774

O'Mahony, M. P., & Smyth, B. (2010). Using readability tests to predict helpful product reviews. In *Adaptivity, personalization and fusion of heterogeneous information* (pp. 164–167). Retrieved January 5, 2024 from https://researchrepository.ucd.ie/server/api/core/bitstreams/7c9bb6a6-34c1-478d-b137-cb77aa5c65c7/content.

Peng, X., & Xu, Q. (2020). Investigating learners' behaviors and discourse content in MOOC course reviews. *Computers & Education, 143*, Article 103673. https://doi.org/10.1016/j.compedu.2019.103673

Pozón-López, I., Higueras-Castillo, E., Muñoz-Leiva, F., & Liébana-Cabanillas, F. J. (2021). Perceived user satisfaction and intention to use massive open online courses (MOOCs). *Journal of Computing in Higher Education, 33*(1), 85–120. https://doi.org/10.1007/s12528-020-09257-9

Qaddumi, H., Bartram, B., & Qashmar, A. L. (2021). Evaluating the impact of ICT on teaching and learning: A study of Palestinian students' and teachers' perceptions. *Education and Information Technologies, 26*(2), 1865–1876. https://doi.org/10.1007/s10639-020-10339-5

Quaderi, S. J. S., & Varathan, K. D. (2024). Identification of significant features and machine learning technique in predicting helpful reviews. *PeerJ Computer Science, 10*, Article e1745. https://doi.org/10.7717/peerj-cs.1745

Ranga, I., Singh, R., & Ranga, B. (2023). Which user-generated content is considered useful by tourists? An investigation into the role of information types shared in online discourse in online travel communities. *International Journal of Human-Computer Interaction, 39*(15), 3114–3126. https://doi.org/10.1080/10447318.2022.2093447

Sandiwarno, S., Niu, Z., & Nyamawe, A. S. (2024). A novel hybrid machine learning model for analyzing e-learning users' satisfaction. *International Journal of Human-Computer Interaction, 40*(16), 4193–4214. https://doi.org/10.1080/10447318.2023.2209986

Schallert, D. L., Sanders, A. J. Z., Williams, K. M., Seo, E., Yu, L.-T., Vogler, J. S., Song, K., Williamson, Z. H., & Knox, M. C. (2015). Does it matter if the teacher is there? A teacher's contribution to emerging patterns of interactions in online classroom discussions. *Computers & Education, 82*, 315–328. https://doi.org/10.1016/j.compedu.2014.11.019

Sebbaq, H., & El Faddouli, N. (2022). Fine-tuned BERT model for large scale and cognitive classification of MOOCs. *International Review of Research in Open and Distributed Learning, 23*(2), 170–190. https://doi.org/10.19173/irrodl.v23i2.6023

Shah, V., Murthy, S., Warriem, J., Sahasrabudhe, S., Banerjee, G., & Iyer, S. (2022). Learner-centric MOOC model: A pedagogical design model towards active learner participation and higher completion rates. *Educational Technology Research & Development, 70*(1), 263–288. https://doi.org/10.1007/s11423-022-10081-4

Shukor, N. A., & Abdullah, Z. (2019). Using learning analytics to improve MOOC instructional design. *International Journal of Emerging Technologies in Learning, 14* (24), 6–17. Retrieved January 5, 2024 from https://www.learntechlib.org/p/217038/.

Soni, S., Chouhan, S. S., & Rathore, S. S. (2023). TextConvoNet: A convolutional neural network based architecture for text classification. *Applied Intelligence, 53*(11), 14249–14268. https://doi.org/10.1007/s10489-022-04221-9

Srivastav, G., Kant, S., & Srivastava, D. (2024). Design of an AI-driven feedback and decision analysis in online learning with Google BERT. *International Journal of Intelligent Systems and Applications in Engineering, 12*(10s), 629–643, 629–643. from https://ijisae.org/index.php/IJISAE/article/view/4465.

Staubitz, T., Klement, H., Renz, J., Teusner, R., & Meinel, C. (2015). Towards practical programming exercises and automated assessment in massive open online courses. In *2015 IEEE international conference on teaching, assessment, and learning for engineering* (pp. 23–30). IEEE. https://doi.org/10.1109/TALE.2015.7386010.

Suzuki, M., Sakaji, H., Hirano, M., & Izumi, K. (2023). Constructing and analyzing domain-specific language model for financial text mining. *Information Processing & Management, 60*(2), Article 103194. https://doi.org/10.1016/j.ipm.2022.103194

Von Helversen, B., Abramczuk, K., Kopeć, W., & Nielek, R. (2018). Influence of consumer reviews on online purchasing decisions in older and younger adults. *Decision Support Systems, 113*, 1–10. https://doi.org/10.1016/j.dss.2018.05.006

Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., & Fu, J. (2023). Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys, 56*(3), 1–52. https://doi.org/10.1145/3611651

Wang, Y., Zhang, J., Yang, Z., Wang, B., Jin, J., & Liu, Y. (2024). Improving extractive summarization with semantic enhancement through topic-injection based BERT model. *Information Processing & Management, 61*(3), Article 103677. https://doi.org/10.1016/j.ipm.2024.103677

Warren, J., Rixner, S., Greiner, J., & Wong, S. (2014). Facilitating human interaction in an online programming course. *Proceedings of the 45th ACM technical symposium on computer science education*. https://doi.org/10.1145/2538862.2538893

Watson, S. L., Watson, W. R., Janakiraman, S., & Richardson, J. (2017). A team of instructors' use of social presence, teaching presence, and attitudinal dissonance strategies: An animal behaviour and welfare MOOC. *International Review of Research in Open and Distributed Learning, 18*(2), 68–91. https://doi.org/10.19173/irrodl.v18i2.2663

Wen, X., & Juan, H. U. (2024). Early prediction of MOOC dropout in self-paced students using deep learning. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2023.2300000

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2023). Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education, 33*(3), 439–466. https://doi.org/10.1007/s40593-022-00290-6

Xue, Z., He, G., Liu, J., Jiang, Z., Zhao, S., & Lu, W. (2023). Re-Examining lexical and semantic attention: Dual-view graph convolutions enhanced BERT for academic paper rating. *Information Processing & Management, 60*(2), Article 103216. https://doi.org/10.1016/j.ipm.2022.103216

Zhao, B., Jin, W., Zhang, Y., Huang, S., & Yang, G. (2023). Prompt learning for metonymy resolution: Enhancing performance with internal prior knowledge of pre-trained language models. *Knowledge-Based Systems, 279*, Article 110928. https://doi.org/10.1016/j.knosys.2023.110928

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications, 41*(4), 1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044

Zhou, C., Yang, S., Chen, Y., Zhou, S., Li, Y., & Qazi, A. (2023). How does topic consistency affect online review helpfulness? The role of review emotional intensity. *Electronic Commerce Research, 23*(4), 2943–2978. https://doi.org/10.1007/s10660-022-09597-x

Zhu, X., Wu, H., & Zhang, L. (2022). Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies, 15*(3), 364–375. https://doi.org/10.1109/TLT.2022.3175537

Zhuang, W., Zeng, Q., Zhang, Y., Liu, C., & Fan, W. (2023). What makes user-generated content more helpful on social media platforms? Insights from creator interactivity perspective. *Information Processing & Management, 60*(2), Article 103201. https://doi.org/10.1016/j.ipm.2022.103201