International Review of Research in Open and Distributed Learning

Fine-Tuned BERT Model for Large Scale and Cognitive Classification of MOOCs

Hanane Sebbaq et Nour-eddine El Faddouli

Volume 23, numéro 2, mai 2022

URI : https://id.erudit.org/iderudit/1089163ar
DOI : https://doi.org/10.19173/irrodl.v23i2.6023

Aller au sommaire du numéro

Éditeur(s)

Athabasca University Press (AU Press)

ISSN

1492-3831 (numérique)

Découvrir la revue

Citer cet article

Sebbaq, H. & El Faddouli, N.-e. (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

Résumé de l'article

The quality assurance of MOOCs focuses on improving their pedagogical quality. However, the tools that allow reflection on and assistance regarding the pedagogical aspects of MOOCs are limited. The pedagogical classification of MOOCs is a difficult task, given the variability of MOOCs' content, structure, and designs. Pedagogical researchers have adopted several approaches to examine these variations and identify the pedagogical models of MOOCs, but these approaches are manual and operate on a small scale. Furthermore, MOOCs do not contain any metadata on their pedagogical aspects. Our objective in this research work was the automatic and large-scale classification of MOOCs based on their learning objectives and Bloom's taxonomy. However, the main challenge of our work was the lack of annotated data. We created a dataset of 2,394 learning objectives. Due to the limited size of our dataset, we adopted transfer learning via bidirectional encoder representations from Transformers (BERT). The contributions of our approach are twofold. First, we automated the pedagogical annotation of MOOCs on a large scale and based on the cognitive levels of Bloom's taxonomy. Second, we fine-tuned BERT via different architectures. In addition to applying a simple softmax classifier, we chose prevalent neural networks long short-term memory (LSTM) and Bi-directional long short-term memory (Bi-LSTM). The results of our experiments showed, on the one hand, that choosing a more complex classifier does not boost the performance of classification. On the other hand, using a model based on dense layers upon BERT in combination with dropout and the rectified linear unit (ReLU) activation function enabled us to reach the highest accuracy value.

International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 23, Number 2

May – 2022

Fine-Tuned BERT Model for Large Scale and Cognitive Classification of MOOCs

Hanane Sebbaq and Nour-eddine El Faddouli
RIME Team, MASI Laboratory, E3S Research Center, Mohammadia School of Engineers (EMI), Mohammed V University, Rabat, Morocco

Abstract

The quality assurance of MOOCs focuses on improving their pedagogical quality. However, the tools that allow reflection on and assistance regarding the pedagogical aspects of MOOCs are limited. The pedagogical classification of MOOCs is a difficult task, given the variability of MOOCs' content, structure, and designs. Pedagogical researchers have adopted several approaches to examine these variations and identify the pedagogical models of MOOCs, but these approaches are manual and operate on a small scale. Furthermore, MOOCs do not contain any metadata on their pedagogical aspects. Our objective in this research work was the automatic and large-scale classification of MOOCs based on their learning objectives and Bloom's taxonomy. However, the main challenge of our work was the lack of annotated data. We created a dataset of 2,394 learning objectives. Due to the limited size of our dataset, we adopted transfer learning via bidirectional encoder representations from Transformers (BERT). The contributions of our approach are twofold. First, we automated the pedagogical annotation of MOOCs on a large scale and based on the cognitive levels of Bloom's taxonomy. Second, we fine-tuned BERT via different architectures. In addition to applying a simple softmax classifier, we chose prevalent neural networks long short-term memory (LSTM) and Bi-directional long short-term memory (Bi-LSTM). The results of our experiments showed, on the one hand, that choosing a more complex classifier does not boost the performance of classification. On the other hand, using a model based on dense layers upon BERT in combination with dropout and the rectified linear unit (ReLU) activation function enabled us to reach the highest accuracy value.

Keywords: cognitive MOOC classification, BERT, LSTM, transfer learning

Introduction

At the end of 2019, the spread of COVID-19 has caused a worldwide change in teaching from face-to-face to virtual or semi-virtual models. To ensure the continuity and efficacy of the learning process, many universities have turned to e-learning and especially MOOCs, and many professors use MOOCs to provide their academic courses. However, there are challenges to overcome, such as course design and quality. Quality, as it relates to the pedagogical framework of e-learning systems is the cornerstone for learning success and effectiveness (Conole, 2016) and it must be subject to constant monitoring and improvement. The quality assurance of e-learning systems can be guaranteed by applying quality instructional design (Kopp & Lackner, 2014) and by analyzing an important aspect of the teaching and learning process, namely the definition of learning objectives (LOs) associated with modules and programs. LOs are central to teaching and learning in many higher education institutions. However, teachers have limited tools to help them reflect on the LOs in the courses they create (Swart & Daneti, 2019). Our research aimed to assist teachers by recommending appropriate content based on their research (Sebbaq & al, 2020) and, importantly, on the cognitive level of their LOs. In our previous work (Sebbaq & Faddouli, 2021), we proposed a pedagogically enriched massive online open course (MOOC) ontology, that served as a standard to unify the representation of MOOCS and facilitate interoperability between MOOCs platforms. We enriched this ontology with metadata about learning objectives classified according to Bloom's taxonomy. This ontology served as a basis for the design and implementation of a linked data repository. We automatically extracted semantically rich descriptive metadata from different MOOC providers and integrate this metadata into a repository accessible through a simple protocol and RDF query language (SPARQL) endpoint. This repository served as the basis for our recommendation framework.

To concretize the epistemological position of the pedagogical dimension of MOOCs, we mapped MOOC learning objectives and Bloom's taxonomy levels. However, one of the limitations of our previous research work has been that the process of MOOC classification remained manual, which was tedious given its large scale. Therefore, the automation of classifying MOOCs according to their cognitive level remained an open research question. In this work, we proposed an approach for the cognitive classification of MOOCs. To the best of our knowledge and according to our literature review, there is no relevant study on automatic and large-scale pedagogical classification of MOOCs.

To study the pedagogies most suited to large-scale learning and teaching, and to highlight the special characteristics and properties of these pedagogies, several studies have compared existing pedagogies and case studies on MOOCs. Those studies analyzed the pedagogies used and proposed mechanisms and guidelines to improve the quality of MOOCs' pedagogical design. Analysis and classification of MOOCS was a difficult task, given the variability of MOOC structures, contents, designs, platforms, providers, and learner profiles. According to our literature review, the pedagogical classification of MOOCs has often been manual, as well as restricted to a limited number of MOOCs whose metadata are extracted manually and on a reduced scale. The objective of our work was to propose an automatic and large-scale pedagogical classification system for MOOCs according to their learning objectives. As we relied on the learning objectives for classification, we adopted the six cognitive levels of Bloom's taxonomy. The main constraint related to our study was the absence of annotated data—there is no dataset of annotated learning objectives organized according to the cognitive levels of Bloom's taxonomy. To overcome this problem, we resorted to building our dataset.

We managed to build a dataset of 2,394 LOs but this size remained limited. The transfer learning technique has demonstrated its performance on small datasets. We proposed a model based on bidirectional encoder representations from Transformers (BERT) for the cognitive classification of MOOCs using various fine-tuning strategies, and we examined the effect of different classifiers upon layers of BERT. Our experiment results showed, on the one hand, that using the pre-trained BERT model and fine-tuning it by adding dense layers outperformed the use of more complex classifiers like long short-term memory (LSTM) or (Bi-LSTM). On the other hand, using dense layers upon BERT in combination with dropout and the rectified linear unit (ReLU) activation function helped avoid overfitting.

The rest of this paper is organized as follows: Section 2 is a literature review. Section 3 describes the methodologies. Section 4 shows the experimental study. Section 5 answers the research questions and Section 6 is a conclusion.

Literature Review

Pedagogical Classification of MOOCs

To improve the quality of the instructional design of MOOCs, several studies have carried out comparisons of pedagogies that are most suited to large-scale learning and teaching, and that highlight the special characteristics and properties of these pedagogies. However, analysis and classification of MOOCs have been a difficult task given the variability of MOOC content, structure, designs, and providers. Educational researchers have adopted several approaches to understand these variations and identify the pedagogical models that exist in the pedagogical design of MOOCs. Kopp and Lackner (2014) studied MOOC models and designs. They structured these elements into a comprehensive checklist in the form of a framework to assist teachers in the design and development of a MOOC. However, this framework was descriptive and did not specify the characteristics associated with either the MOOC dimensions or assessment. Yousef et al. (2014) conducted research that classified MOOC quality criteria in two dimensions and six categories, which were manifested via 74 criteria. However, this study did not go further to evaluate MOOCs based on these criteria.

After a review of classification and description systems of existing MOOCs, Major and Blackmon (2016) proposed a descriptive framework with 11 dimensions including the educational dimension. Nevertheless, they did not go as far as pedagogical assessment. Similarly, to describe MOOCs, Rosselle et al. (2014) mapped eight dimensions to various characteristics of MOOCs. However, no assessment system was proposed. This mapping was an extension of that proposed by Pardos and Schneider (2013) who provided a conceptual mapping of MOOC designs. They categorized five main dimensions, which included four elements of the learning environment that could potentially affect design—instruction, content, assessment, and community.

To provide teachers with the guidance and the assistance they need to make better design decisions, Conole (2014) offered the 7Cs of learning design framework. This framework can be used for both designing and evaluating MOOCs. Moreover, Conole (2016) has also offered a 12-dimensional framework as well as a

rating scale (low, medium, or high). These dimensions covered structural, philosophical, and pedagogical aspects, though the organizational system can be confusing. In addition, the assessment of some dimensions was unclear. Margaryan et al. (2015) proposed the course scan assessment system, a 37-item checklist based on existing instruments for quality instructional design. Margaryan et al. evaluated a sample of 76 MOOCs using the three dimensions: (a) course details (7 elements); (b) objectives and organization (6 elements); and (c) pedagogical principles (24 elements). The MOOCs evaluated were a random sample of those available at the end of 2013 on various platforms.

Based on an approach focused on pedagogy Swan et al. (2014) offered the MOOC assessing MOOC pedagogy (AMP) tool for evaluating pedagogy. The AMP generates a specific MOOC profile based on 10 pedagogical dimensions: (a) epistemology, (b) role of the teacher, (c) orientation of activities, (d) structure, (e) approach to content, (f) feedback, (g) cooperative learning, (h) adaptation to individual differences, (i) activities and evaluation, and (j) user's roles. The rating scale ranged from 0 to 5. Quintana and Tan (2019) introduced an extended version of the AMP tool with modified terminology and more sophisticated indicators. After evaluating 20 MOOCs (from the same platform and institution, but different fields), they showed how machine learning with the k-nearest neighbor (k-NN) algorithm helped identify pedagogically similar MOOCs. Xing (2019), using machine learning for the classification of MOOCs, analyzed 205 MOOCs to identify clusters of MOOCs using the k-means algorithm. Their goal was to study the impact of design features on learner engagement. Davis et al. (2018) used hierarchical clustering to group MOOCs according to their structures. They manually collected data from 177 MOOCs and looked only at the MOOCs' structures. An automatic notation was made by calculating the similarities via the two approaches (clustering transition probability and trajectory mining).

Table 1 summarizes the works reviewed here and classifies them according to whether their objective was description or evaluation. Information on the number of MOOCs analyzed is also provided to assess the large-scale character of the classification, while the data gathering column indicates whether the study gathered data automatically or manually. The third point of comparison shows whether the work integrated a MOOC assessment tool and whether it was automatic or manual. The sixth aspect of comparison deals with the use of machine learning for automating the classification, and the last column addresses the use of a theoretical foundation.

Table 1

Comparing Studies That Classify MOOCs

Research paper	Research objective	Number of MOOCs analyzed	Data gathering method	Assessment tool	Assessment method	Machine learning	Theoretical foundation
Conole (2014)	Description	-	-	Yes (low, medium, high)	Manual	-	Good learning
Conole (2016)	Description Evaluation (7Cs of learning design)	-	-	Yes	Manual	-	Good learning principles

Major and Blackmon (2016)	Description	-	-	-	-	-	-
Pardos and Schneider (2013)	Description Evaluation		-	Yes	Manual	-	Online learning
Rosselle et al. (2014)	Description	-	-	-	-	-	-
Kopp and Lackner (2014)	Description	-	-	-	-	-	ADDIE, Molenda (2003)
Yousef et al. (2014)	Description	-	-	-	-	-	-
Swan et al. (2014)	Pedagogical evaluation	17	Manual	AMP tool	Manual	-	-
Margaryan et al. (2015)	Evaluation	76	Manual	Course scan	Manual	-	Merrill
Xing (2019)	Evaluation	205	Manual	Yes	Manual	K-means	Web-based online instruction
Davis et al. (2018)	Evaluation	177	Manual	Calculate similarities	Automatic	Hierarchical clustering	-
Quintana and Tan (2019)	Evaluation (AMP tool extension)	20	Manual	Expanded AMP instrument	Manual	k-NN	-

Pedagogical Classification of E-Learning Content Based on Bloom's Taxonomy

Classification based on Bloom's taxonomy makes use of Bloom's taxonomy action verbs (BTAV). Swart and Daneti (2019) and Nevid and McClelland (2013) used BTAV to manually classify LOs. This time-consuming task required the participation of educational specialists in order to assure accurate outcomes. On the other hand, some verbs in the BTAV are unclear, since determining the necessary cognitive level is challenging. Verbs such as choose, describe, design, explain, show, and use can be found at different levels of cognition.

The use of Bloom's taxonomy to classify e-learning content (e.g., questions, forum texts) has received considerable attention in recent years. For automatic classification, researchers have used a variety of methodologies, ranging from rule-based to traditional machine learning to deep learning approaches. In the 1980s, the rule-based approach was the most popular. Omar (2012) and Haris and Omar (2012) demonstrated the effectiveness of the rule-based approach, although it was not without flaws. Among these disadvantages was the requirement for professionals to manually construct many rules to cover all sorts and domains of inquiries in order to improve the accuracy of the output.

When it comes to e-learning content, researchers have been mainly interested in evaluation questions. Abduljabbar and Omar (2015) combined three classifiers—support vector machine (SVM), nearest

neighbor (NB), and k-nearest neighbor (k-NN)—and found that using an only bag of words (BOW) feature extraction improved accuracy. In 2020, Mohammed and Omar (2020) used a combination of Term Frequency–Inverse Document Frequency (TF-IDF), Part Of Speech (POS), and word2vec for feature extraction and tested the classifiers (i.e., k-NN, logistic regression (LR), and SVM). This combination, according to their research, increased F1-measurement performance. Osman and Yahya (2016) combined multiple feature extraction (BOW, POS, and n-grams) techniques with different machine learning algorithms (i.e., NB, SVM, LR, and decision trees) in order to test and compare them. Their research revealed that the feature extraction technique used had an impact on the machine learning classifier's performance.

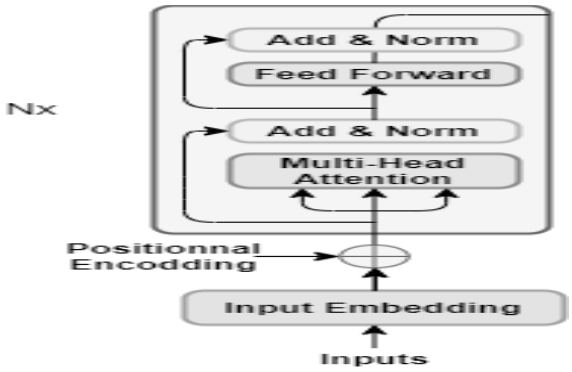
Some studies have used a deep learning-based approach to classify e-learning content according to the six cognitive levels of Bloom's Taxonomy. Ting Fei et al. (2003) examined the application of automated question classification tests in e-learning systems. They presented a text classification model that used a back-propagation learning approach to train a text classifier using an artificial neural network. Their technology outperformed the competition by about 78% in terms of F1 value. Yusof and Hui (2010) used an artificial neural network (ANN) strategy that employed numerous feature reduction strategies to develop a model that categorized question items. Das et al. (2020) proposed two strategies for automatically estimating the cognitive learning challenges of given questions. Their first method used latent Dirichlet allocation (LDA) as a deep learning strategy. For multi-class text classification, the second methodology employed BERT. According to their findings, BERT had an accuracy of 89%, which was higher than LDA's 83%.

BERT for Text Classification

The BERT model is based on two stages: pre-training and fine-tuning (Devlin et al., 2019). During pre-training, the model is trained on a large unlabeled corpus. The model is then fine-tuned, starting with the pre-trained parameters and refining all parameters with task-specific labeled data. BERT uses the transformer that is a new architecture presented in Vaswani et al. (2017). A simple transformer consists of an encoder that reads text input and a decoder that generates a task prediction. BERT requires only the encoder depicted in Figure 1 because its objective is to develop a model of the language representation.

Figure 1

The BERT Encoder



BERT is based on the attention mechanism (Vaswani et al., 2017) that was invented to allow a model to comprehend and remember the contextual relationships between features and text. The attention mechanism maps a set of queries to their corresponding sets of keys and values—vectors that contain information about related or neighboring entities from the text input. The procedure involves the dot product of the input query with the existing keys, and then a softmax to give a scaled dot product attention score, which is defined as follows (Vaswani et al., 2017):

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

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the Q and K matrices. This resulting score vector is then multiplied by each value and summed to give the final self-attention result for that particular query. Interpretation of this multi-head attention helps the model determine how much attention it should pay to each word in the text block (Vaswani et al., 2017). Multi-head attention is defined as follows:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O$$

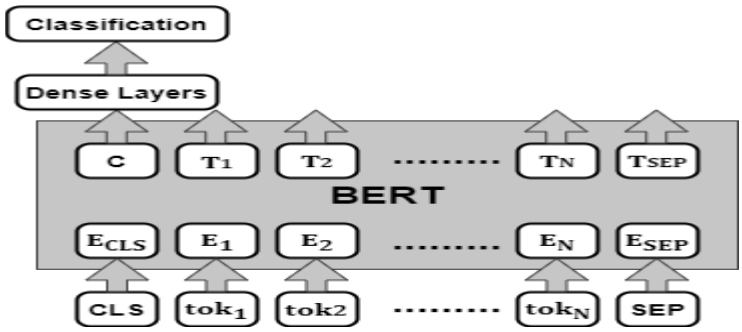
where $head_i$ = Attention (QW^Q_i, KW^K_i, $V W^V_i$) and the projections are parameter matrices $W_i^Q \in \mathbb{R}^{dmodel \times dk}$, $W_i^K \in \mathbb{R}^{dmodel \times dk}$, $W_i^V \in \mathbb{R}^{dmodel \times dv}$, and $W^O \in \mathbb{R}^{hdv \times dmodel}$ (Vaswani et al., 2017).

BERT represents a single sentence or a pair of sentences as a sequence of tokens with the following characteristics:

- The first token in the sequence is [CLS].
- When there is a pair of sentences in the sequence, they are separated by the token [SEP].
- For a given token, its input representation is constructed by summing the corresponding token, position, and segment embeddings (see Figure 2).

Figure 2

BERT Architecture



BERT is a leading model for a variety of Natural Language Processing (NLP) tasks, demonstrating its efficiency and potential. In this study, we explored fine-tuning methods for applying BERT to a cognitive classification task.

Methodology

Theoretical Foundation of Our Proposed Approach

According to our comparative study summarized in Table 1, several theoretical frameworks have been adopted for the evaluation and classification of MOOCs. Conole (2014) associated good learning with quality learning. It was critical, in his opinion, to meet the characteristics of good learning in order to accomplish effective learning. Conole (2016) based the12-dimensional assessment framework, as well as the 7Cs for learning design framework on this principle. On the other hand, according to Merrill (2012), the first principles of instruction he proposed constituted the foundation of all present pedagogical models and theories. Merill suggested five guidelines for the development of learning activities. Margaryan et al. (2015) built on Merill's first principles of instruction and added five more principles related to learning resources. Conole (2014, 2016) and Magrayan et al. (2015) used these two theoretical frameworks to drive their research into the development of evaluation frameworks focused on open-ended questions and necessitating the assistance of an expert. Xing (2019) used a Web-based online instruction approach to drive their evaluation of MOOCs. This approach was more generic and incorporated three global design dimensions: information, instruction, and learning. All the frameworks adopted in these studies focused on open-ended questions and called for expert assistance, while our research objective was to automate the evaluation of LOs.

The ultimate aim of our research was to evaluate and classify the pedagogical dimension of MOOCs based on their learning objectives. The theoretical foundation of Bloom's taxonomy was most appropriate for our context since it covered the different levels of cognitive learning and allowed for classifying learning objectives according to six hierarchical levels. The initial purpose of Bloom's taxonomy was to assist teachers in developing rubrics and measuring the achievement of their learning goals by providing

guidelines. We used a modification of Bloom's taxonomy adapted from Krathwohl (2002) who proposed a revision of the original taxonomy. It defined a two-dimensional framework consisting of knowledge and cognitive processes. The first dimension took the subcategories of the first level of the original taxonomy; the second dimension renamed the six levels as verbs—remembering, understanding, applying, analyzing, evaluating, and creating. Our research considered the dimension of cognitive learning processes.

Most studies that have classified e-learning content based on Bloom's taxonomy focused on classifying assessment questions. No research has been done on the automatic classification of LOs. Machine learning has been used most often, followed by the rule-based approach. The majority of research publications have focused on merging multiple feature extraction and feature selection methods to improve the performance of machine learning classifiers. The deep learning approach has been used less often; only the ANN architecture has been tested in this context. BERT was used in a single study for cognitive classification purposes (Das et al., 2020). There has been some research comparing BERT and other machine learning or deep learning models (González-Carvajal & Garrido-Merchán, 2020). Our research, on the other hand, is the first to investigate the cognitive classification of LOs.

A BERT-Based Cognitive Approach for Classifying MOOCs

From our review of the literature, we have deduced that there is no large-scale, automatic classification system for MOOCs based on their pedagogical approaches. As we summarized in Table 1, the existing research has addressed one of the following:

- Frameworks developed for quality assurance that are generalist and lack educational considerations and means to operationalize the review of MOOCs' pedagogical quality.
- Case studies that detailed the design of individual MOOCs to highlight best practices and pedagogical models. However, these studies concerned only a small number of MOOCs and were not based on a well-defined evaluation framework.
- Descriptive frameworks that were intended for designing MOOCs from scratch, which was not our objective.
- Evaluation frameworks that dealt with several dimensions including the pedagogical.

However, not all frameworks of the latter type were focused on pedagogy, and they all suffered from the lack of an automatic and large-scale system for classifying MOOCs according to their pedagogical models. In addition, their dimensions were broad and had to be operationalized via qualitative and quantitative indicators as well as concrete characteristics. This research analyzed course design and pedagogy to understand variations in the two, but much of this analysis relied on a human categorization process based on broad interpretations of the learning designs. In addition, Assessment tools were based on open-ended questions that required the intervention of an expert. Since it is difficult to automate the assessment, this has remained a manual task and automated tools are not yet widely adopted by researchers in the MOOC community. The only study whose assessment was automatic, Davis et al. (2018) was restricted to comparing MOOC structures, similar to Pardos and Schneider (2013). Our objective was to classify MOOCs on a large scale; the number of MOOCs analyzed in the research cited above was not sufficient to deduce

the different pedagogies in MOOCs. Both Xing (2019) and Davis et al. (2018) used machine learning for a large analysis of MOOCs. However, the number of MOOCs they examined remained limited, and their data collection methods were manual. Swan et al. (2014) used machine learning for the analysis of about 20 MOOCs. Nevertheless, the result of their clustering cannot be generalized given the limited number of MOOCs they analyzed.

The main challenge of our study was the lack of annotated data. Despite thorough research, we found no annotated learning goal datasets, so we created our learning objectives dataset. Evens so, the size of the dataset remained limited. For its part, BERT is the state-of-the-art technique in NLP (Devlin et al., 2018) and it has demonstrated its performance on small datasets. The contributions of this study are:

- The automatic classification of MOOCs according to their pedagogical approaches based on the cognitive levels of Bloom's taxonomy. The first phase of our automated approach has been implemented in our previous work (Sebbaq & Faddouli, 2021).
- The large scale of our approach was based on the repository of MOOCs already built in our previous work (Sebbaq & Faddouli, 2021).
- The use of transfer learning to resolve the issue of the lack of annotated data.
- Fine-tuning BERT via different strategies: we investigated the impact of choosing different classifiers, from a simple softmax classifier to a more complex classifier like dense layers, LSTM, and Bi-LSTM.

Fine-Tuning Strategies

BERT fine-tuning involves training a classifier with different layers on top of the pre-trained BERT transformer to minimize task-specific parameters. Fine-tuning for a specific task can be done using several approaches, either by fine-tuning the architecture or by fine-tuning different hyper-parameters such as the learning rate or the choice of the best optimization algorithm. Our objective in this research work was the cognitive classification of MOOCs according to their learning objectives. For this type of classification problem, we simply adopted the basic architecture of BERT and then added an output layer for the classification. This layer took as input the final hidden state of the first token [CLS]. We considered this exit to be the ultimate representation of the entire entry sequence. The output layer can be either a simple classifier like softmax or a more complicated network like the bidirectional Bi-LSTM. In our approach, we proposed six different architectures for fine-tuning BERT.

BERT-Based Fine-Tuning

Figure 3 represents the first architecture. In this basic architecture, we mainly relied on the BERT base; we used the output of the token [CLS] provided by BERT only. The output [CLS] was a vector of size 768; we gave it as input to a network that was fully connected with no hidden layer. Since our classification problem was multi-class, the output layer was based on a softmax activation layer. The softmax function formula is:

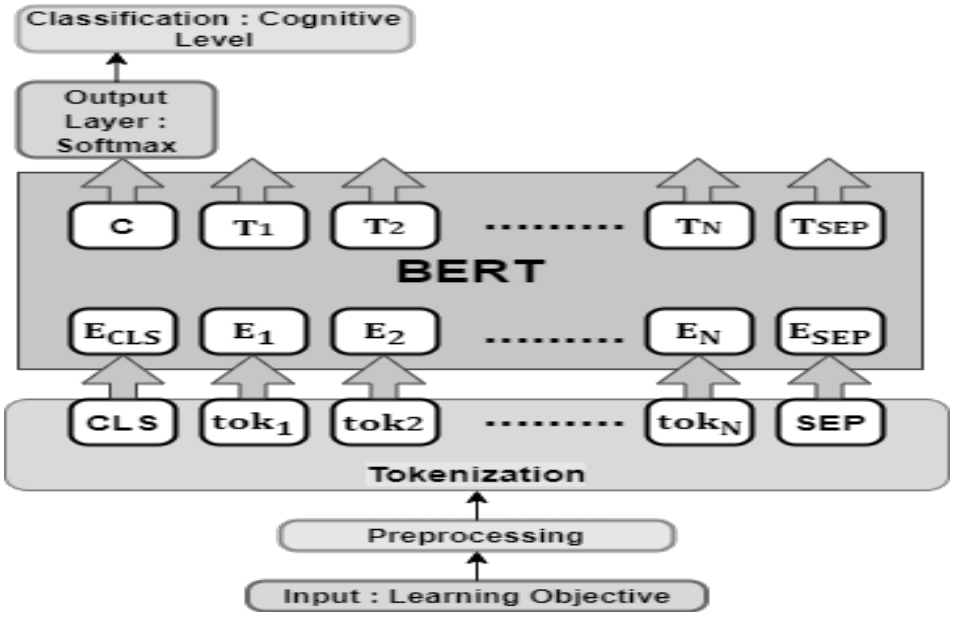
$$\sigma(\vec{z}) = \frac{e^{\mathrm{zi}}}{\sum_{j=1}^{K} e^{zj}}$$

where $\vec{Z} = (z_1; : : : ; z_K)$, z_i values are the elements of the input vector to the softmax function, K is the number of classes in the multi-class classifier. The output node with the highest probability is then chosen as the predicted label for the input.

For preprocessing, we simply cleaned the text of non-alphabetic characters and converted it to lower case.

Figure 3

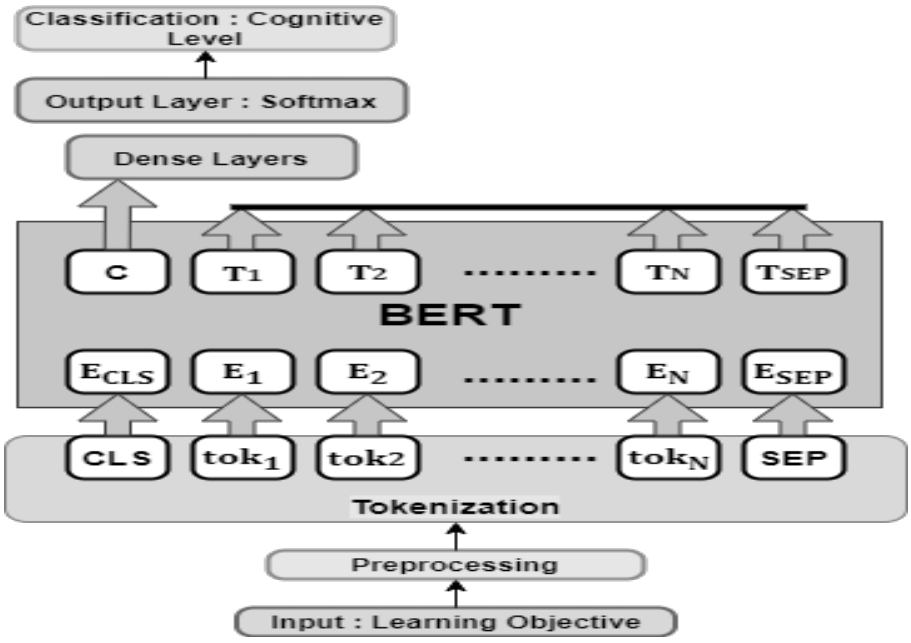
BERT-Based Fine-Tuning Architecture



In this architecture (see Figure 4), we added fully connected layers. The fully connected layer took the output of BERT's 12 layers and transformed it into the final output of six classes that represented the six cognitive levels of Bloom's taxonomy. This layer consists of three dense layers.

Figure 4

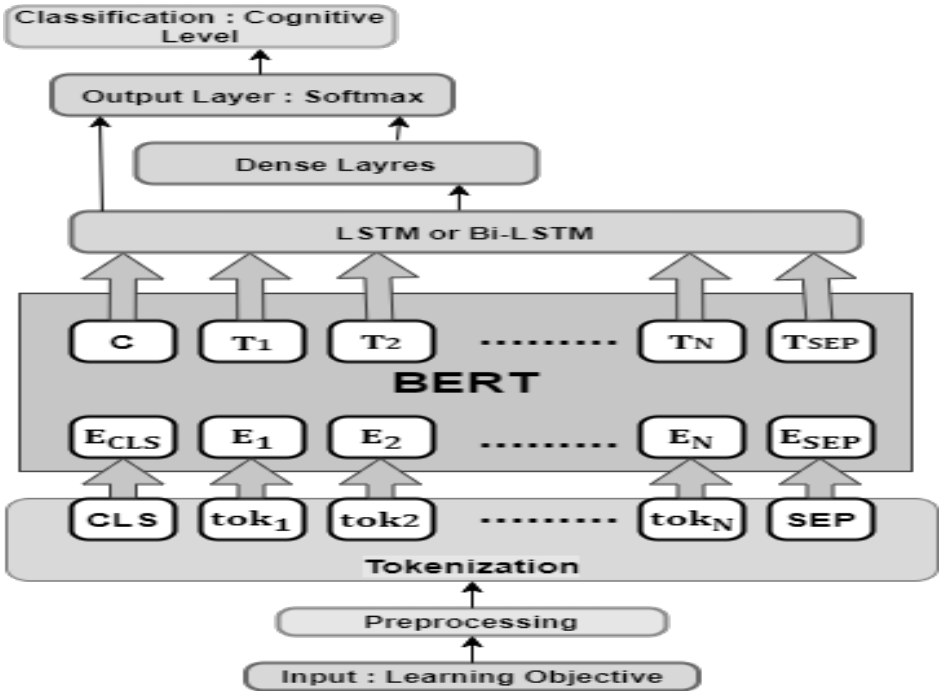
BERT With Fully Connected Layers Architecture



In previous architectures, the [CLS] output was the only one used as input for the classifier. In this architecture, we used all the outputs of the last transformer encoder as inputs to an LSTM or Bi-LSTM recurrent neural network as shown in Figure 5. After the input was processed, the network sent the final hidden state to the output layer, which was a fully connected network, to perform classification using the softmax activation function. We also experimented with architecture that was more complex by inserting dense layers between the deep network layer and the output layer.

Figure 5

BERT With Fully Connected Layers and Deep Network Layers Architecture



Experimental Study

In this section, we provide a representation and a detailed analysis of the dataset as well as a complete presentation of the results that we obtained from experimentation with the different models.

Dataset Description and Analysis

Given the challenge posed by the lack of annotated datasets of LOs according to the cognitive levels of Bloom's taxonomy, our solution was to create our dataset as presented in Table 2. We started by collecting LOs from the MOOCs providers, Coursera, and edX, and then manually annotated them based on Bloom's taxonomy action verbs list. However, some of the action verbs in BTAV overlap at several levels of the hierarchy (Krathwohl, 2002). This leads to ambiguity about the actual meaning of the required cognition (Stanny, 2016) and affects the effectiveness of the BTAV-based classification. Moreover, this method has the drawback of not being able to guarantee the accuracy of our annotations, as well as being time-consuming.

Table 2

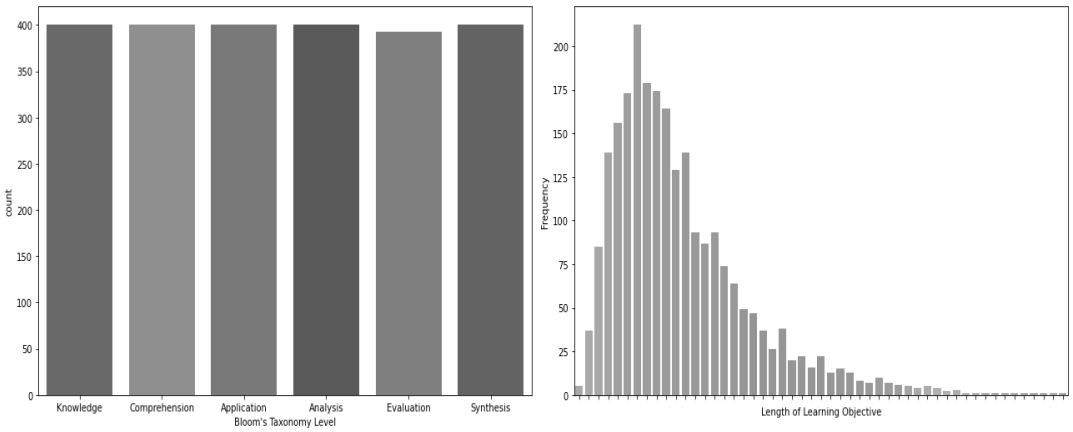
The Distribution of LOs in Our Dataset

Cognitive level	2 394	Example
Knowledge (remembering)	400	Describe the concept of modular programming and the uses of the function in computer programming
Comprehension (understanding)	400	At the end of this module, the learner will be able to classify clustering algorithms based on the data and cluster requirements
Application (applying)	400	Apply a design process to solve object-oriented design problems
Analysis (analyzing)	400	Analyze the appropriate quantization algorithm
Evaluation (evaluating)	394	Compare the semantic and syntactic ways encapsulation
Synthesis (creating)	400	Create a Docker container in which you will implement a Web application by using a flask in a Linux environment

The training dataset consisted of 2,394 training objectives. Figure 6 illustrates the number of words per input data point in the form of a histogram. According to the histogram, the average length of the training objectives was about 225. Regarding the class distribution of the data in the input dataset, analysis of Figure 6 suggests that the classes are balanced, with 400 learning objectives per cognitive level.

Figure 6

Distribution of Data by Level and Item Length



Evaluation Metrics

Several considerations, including class balance and expected outcomes, guided the selection of the best measures to evaluate the performance of a given classifier on a certain dataset. Given a dataset with an

approximately balanced number of samples from all classes, we used the accuracy measure to evaluate the performance of our model and compare it with other models (Grandini et al., 2020). Accuracy is the sum of true positive (TP) and true negative (TN) items divided by the sum of all other possibilities, defined as follows:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

where TP = True Positives, TN = True Negatives, FN = False Negatives, and FP = False Positives.

Environment Setup

We used the Google Colab and Tensorflow environment as well as Keras Tensorflow to build the BERT models. Keras TensorFlow is an open-source mathematical software library used for machine learning applications. It has tools to run on graphic processing units, which can significantly reduce training and inference times on some models. Keras is a high-level API for TensorFlow. It has a modular and easily extensible architecture, and it allows users to create sequential models or a graph of modules that can be easily combined. The library contains many different types of neural layers, cost, and activation functions. We implemented different fine-tuning strategies of BERT on Tensorflow Hub (TFHub). TFHub provides a way to try, test, and reuse machine learning models.

Implementation Details

For our experiments, we used the basic pre-trained model bert_multi_cased_L-12_H-768_A-12/2, which had 12 layers, 768 hidden, 12 self-attention heads, and 110M parameters. We used the Adam optimizer, which is one of the most stable and widely used in the deep learning world (Kingma & Ba, 2014). Adam was used in combination with the warmup steps, which were low learning-rate updates that helped the model converge. After trying many different configurations, and after numerous unsuccessful attempts that ran out of memory, we arrived at a working configuration of hyper-parameters. The base learning rate was 3e-5, and the warm-up proportion was 0.1. We empirically set the maximum number of epochs to 15 with a batch size of 32 and saved the best model on the validation set for testing.

For the implementation of the models adopted, we use the Keras Layer function of Tensorflow Hub to build our BERT layer. Then we tokenized our text based on the variables of this layer. This allowed us to have the first input of our BERT model, which was input_word_ids. Then we built the two other inputs of BERT, which were the embeddings of position input_mask and segments segment_ids. We added a dropout of 0.1 after each layer.

Results Analysis

We conducted experiments to demonstrate the efficiency of our proposed approach in terms of performance. Our main task was to explore the performance of BERT on cognitive text classification and evaluate the impact of different fine-tuning strategies. We used six models: (a) standard BERT-based fine-tuning, (b) BERT with fully connected layers, (c) BERT with LSTM, (d) BERT with Bi-LSTM, (e) BERT with both LSTM and fully connected layers, and (f) BERT with both Bi-LSTM and fully connected layers.

In particular, we aimed to answer the following research questions via our experiments.

- (RQ1): How do different fine-tuning strategies have different impacts on the cognitive classification task?

- (RQ2): How effectively does our BERT-based cognitive approach produce better results than other baseline fine-tuning strategies?

(RQ1) Comparing Performance of Various Architectures With Different Classifiers

In order to answer (RQ1), we investigated the effects of various classifiers on BERT. To use a basic softmax classifier upon the last layer of BERT, we experimented with a cascade of dense layers with the activation function ReLU and the more complex classifiers, LSTM and Bi-LSTM. Table 3 presents the accuracies of the six models ranked from the basic to the more complicated. The results demonstrate that the use of a more complex classifier did not improve performance. Instead, it lowered accuracy on the five classification models, which is understandable given that BERT also has deep networks and advanced training techniques.

(RQ2) BERT With Three Dense Layers Performed Best

As a response to (RQ2), our proposed model performs better than other baseline fine-tuning strategies. Thanks to the addition of dense layers on top of BERT. A dense layer is a regular, deeply connected neural network layer of neurons. Each neuron in the previous layer receives feedback from all the neurons in the layer before it, making it densely connected. To prevent overfitting, we also used the dropout, a regularization technique where randomly selected neurons are ignored during training. At each upgrade of the training process, dropout randomly set the outgoing edges of hidden units to zero at random. We added a dropout of 0.1 after each dense layer. We used the activation function ReLU. The biggest benefit of using the ReLU mechanism over other activation functions was that it did not simultaneously stimulate any of the neurons.

Table 3

Accuracies of the Different Models

BERT Model	Accuracy
BERT-based fine-tuning	88.75%
BERT with three fully connected layers	92.5%
BERT with LSTM	91.25%
BERT with Bi-LSTM	90.83%
BERT with three fully connected layers + LSTM	91.46%
BERT with three fully connected layers + Bi-LSTM	92.08%

Conclusion

The main constraint of our study was the availability of annotated LOs datasets. We managed to build a dataset of 2,394 LOs but this size was still limited. On the other hand, each LO needed to be carefully annotated for the training data to be correct. This makes building a larger dataset a cumbersome and difficult task to handle.

In this study, our goal was to propose a model for the automatic classification of MOOCs according to their pedagogical approaches, and this on a large scale, based on the cognitive levels of Bloom's taxonomy. To this end, we opted for BERT, and then we experimented with different strategies to fine-tune it. In this sense, we investigated the impact of choosing different classifiers upon BERT, from a simple softmax classifier to a more complex classifier such as dense layers, LSTM, and Bi-LSTM. The results demonstrated that using a more complex classifier did not improve performance. Instead, it lowered accuracy on the five classification models, which is understandable given that BERT also has deep networks and advanced training techniques. We also demonstrated that the use of dense layers upon BERT in combination with dropout and the activation function ReLU allowed us to reach the highest accuracy value. Although BERT with dense layers performed well in our experiment, we have not yet explored other fine-tuning strategies. In a future study, we will tackle other techniques such as multitask learning to enhance the performance of our BERT model.

Overall, our proposed approach proved its ability to classify learning objectives in MOOCs. Since our approach was based on learning objectives for the pedagogical classification of MOOCs, potential applications to other learning objects in the context of distance learning are worth exploring in future research and practice.

References

Abduljabbar, D. A., & Omar, N. (2015). Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination. *Journal of Theoretical and Applied Information Technology, 78*(3), 447. http://www.jatit.org/volumes/Vol78No3/15Vol78No3.pdf

Conole, G. (2014, April). The 7Cs of learning design: A new approach to rethinking design practice. *Proceedings of the Ninth International Conference on Networked Learning* (pp. 502–509). Edinburgh, Scotland. https://www.lancaster.ac.uk/fss/organisations/netlc/past/nlc2014/abstracts/pdf/conole.pdf

Conole, G. (2016). MOOCs as disruptive technologies: Strategies for enhancing the learner experience and quality of MOOCs. *Revista de Educación a Distancia, 50*(2). http://dx.doi.org/10.6018/red/50/2

Das, S., Das Mandal, S. K., & Basu, A. (2020). Identification of cognitive learning complexity of assessment questions using multi-class text classification. *Contemporary Educational Technology, 12*(2), Article ep275. https://doi.org/10.30935/cedtech/8341

Davis, D., Seaton, D., Hauff, C., & Houben, G. J. (2018, June). Toward large-scale learning design: Categorizing course designs in service of supporting learning outcomes. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 1–10). https://doi.org/10.1145/3231644.3231663

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

González-Carvajal, S., & Garrido-Merchán, E. C. (2020). *Comparing BERT against traditional machine learning text classification*. arXiv preprint arXiv:2005.13012. https://arxiv.org/abs/2005.13012

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for multi-class classification: An overview*. arXiv preprint arXiv:2008.05756. https://arxiv.org/abs/2008.05756

Haris, S. S., & Omar, N. (2012, December). A rule-based approach in Bloom's taxonomy question classification through natural language processing. *Seventh International Conference on Computing and Convergence Technology* (ICCCT; pp. 410–414). Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/abstract/document/6530368

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980. https://arxiv.org/abs/1412.6980

Kopp, M., & Lackner, E. (2014). Do MOOCs need a special instructional design? *Proceedings of Sixth International Conference on Education and New Learning* (EDULEARN14; pp. 7138–7147). Barcelona, Spain. https://library.iated.org/view/KOPP2014DOM

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Major, C. H., & Blackmon, S. J. (2016). Massive open online courses: Variations on a new instructional form. *New Directions for Institutional Research, 2015*(167), 11–25. https://doi.org/10.1002/ir.20151

Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). *Computers & Education, 80*, 77–83. https://doi.org/10.1016/j.compedu.2014.08.005

Merrill, M. D. (2012). *First principles of instruction*. John Wiley & Sons. https://digitalcommons.usu.edu/usufaculty_monographs/100/

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLOS ONE 15*, e0230442. https://doi.org/10.1371/journal.pone.0230442

Molenda, M. (2003). In search of the elusive ADDIE model. *Performance improvement*, *42*(5), 34-37. http://www.damiantgordon.com/Courses/DT580/In-Search-of-Elusive-ADDIE.pdf

Nevid, J. S., & McClelland, N. (2013). Using action verbs as learning outcomes: Applying Bloom's taxonomy in measuring instructional objectives in introductory psychology. *Journal of Education and Training Studies, 1*(2), 19–24. https://doi.org/10.11114/jets.v1i2.94

Omar, N., Haris, S. S., Hassan, R., Arshad, H., Rahmat, M., Zainal, N. F. A., & Zulkifli, R. (2012). Automated analysis of exam questions according to Bloom's taxonomy. *Procedia: Social and Behavioral Sciences, 59*, 297–303. https://doi.org/10.1016/j.sbspro.2012.09.278

Osman, A., & Yahya, A. A. (2016). Classifications of exam questions using linguistically-motivated features: A case study based on Bloom's taxonomy. https://www.researchgate.net/publication/298286164_Classifications_of_Exam_Questions_Using_Linguistically-Motivated_Features_A_Case_Study_Based_on_Blooms_Taxonomy

Pardos, Z. A., & Schneider, E. (2013). *AIED 2013 workshops proceedings* (Vol. 1). http://people.cs.pitt.edu/~falakmasir/docs/AIED2013.pdf

Quintana, R. M., & Tan, Y. (2019). Characterizing MOOC pedagogies: Exploring tools and methods for learning designers and researchers. *Online Learning, 23*(4), 62–84. https://doi.org/10.24059/olj.v23i4.2084

Rosselle, M., Caron, P. A., & Heutte, J. (2014, February). A typology and dimensions of a description framework for MOOCs. In *Proceedings of European MOOCs Stakeholders Summit 2014,* (eMOOCs 2014; pp. 130–139). Lausanne, Switzerland. Proceedings document published by Open Education Europa (www. openeducationeuropa. eu). https://hal.archives-ouvertes.fr/hal-00957025/

Sebbaq, H., El Faddouli, N. E., & Bennani, S. (2020, September). Recommender system to support MOOCs teachers: Framework based on ontology and linked data. *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications,*Article 18. https://doi.org/10.1145/3419604.3419619

Sebbaq, H., & Faddouli, N. E. E. (2021, January). MOOCs semantic interoperability: Towards unified and pedagogically enriched model for building a linked data repository. *International Conference on Digital Technologies and Applications*. Springer, 621-631.

Stanny, C. J. (2016). Reevaluating Bloom's taxonomy: What measurable verbs can and cannot say about student learning. *Education Sciences, 6*(4), 37. https://doi.org/10.3390/educsci6040037

Swan, K., Day, S., Bogle, L., & van Prooyen, T. (2014). AMP: A tool for characterizing the pedagogical approaches of MOOCs. *E-mentor, 2*(54), 75–85. https://doi.org/10.15219/em54.1098

Swart, A. J., & Daneti, M. (2019, April). Analyzing learning outcomes for electronic fundamentals using Bloom's taxonomy. *2019 IEEE Global Engineering Education Conference* (EDUCON; pp. 39–44). Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/document/8725137

Ting Fei, Wei Jyh Heng, Kim Chuan Toh, & Tian Qi. (2003). Question classification for e-learning by artificial neural network. *Proceedings of the Joint Fourth International Conference on Information, Communications and Signal Processing, and the Fourth Pacific Rim Conference on Multimedia* (pp. 1757–1761). Institute of Electrical and Electronics Engineers. Singapore. https://doi.org/10.1109/ICICS.2003.1292768

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems* (NIPS 2017; pp. 5998–6008). Long Beach, USA. https://arxiv.org/abs/1706.03762

Xing, W. (2019). Exploring the influences of MOOC design features on student performance and persistence. *Distance Education, 40*(1), 98–113. https://doi.org/10.1080/01587919.2018.1553560

Yousef, A. M. F., Chatti, M. A., Schroeder, U., & Wosnitza, M. (2014, July). What drives a successful MOOC? An empirical examination of criteria to assure design quality of MOOCs. *14th International Conference on Advanced Learning Technologies* (pp. 44–48). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICALT.2014.23

Yusof, N., Hui, C. J. (2010). Determination of Bloom's cognitive level of question items using artificial neural network. *10th International Conference on Intelligent Systems Design and Applications* (ISDA; pp. 866–870). Institute of Electrical and Electronics Engineers. Cairo, Egypt. https://doi.org/10.1109/ISDA.2010.5687152


Athabasca
University


BY