Automatic Classification of Learning Objectives Based on Bloom's Taxonomy

Yuheng Li
Centre for Learning Analytics
Monash University, Australia
yuheng.li@monash.edu

Mladen Raković
Centre for Learning Analytics
Monash University, Australia
mladen.rakovic@monash.edu

Boon Xin Poh
Monash University, Australia
pohjessica96@hotmail.com

Dragan Gašević
Centre for Learning Analytics
Monash University, Australia
dragan.gasevic@monash.edu

Guanliang Chen*
Centre for Learning Analytics
Monash University, Australia
guanliang.chen@monash.edu

ABSTRACT

Learning objectives, especially those well defined by applying Bloom's taxonomy for Cognitive Objectives, have been widely recognized as important in various teaching and learning practices. However, many educators have difficulties developing learning objectives appropriate to the levels in Bloom's taxonomy, as they need to consider the progression of learners' skills with learning content as well as dependencies between different learning objectives. To remedy this challenge, we aimed to apply state-of-the-art computational techniques to automate the classification of learning objectives based on Bloom's taxonomy. Specifically, we collected 21,380 learning objectives from 5,558 different courses at an Australian university and manually labeled them according to the six cognitive levels of Bloom's taxonomy. Based on the labeled dataset, we applied five conventional machine learning approaches (i.e., naive Bayes, logistic regression, support vector machine, random forest, and XGBoost) and one deep learning approach based on pre-trained language model BERT to construct classifiers to automatically determine a learning objective's cognitive levels. In particular, we adopted and compared two methods in constructing the classifiers, i.e., constructing multiple binary classifiers (one for each cognitive level in Bloom's taxonomy) and constructing only one multi-class multi-label classifier to simultaneously identify all the corresponding cognitive levels. Through extensive evaluations, we demonstrated that: (i) BERT-based classifiers outperformed the others in all cognitive levels (Cohen's κ up to 0.93 and F1 score up to 0.95); (ii) three machine learning models – support vector machine, random forest, and XGBoost — delivered performance comparable to the BERT-based classifiers; and (iii) most of the binary BERT-based classifiers (5 out of 6) slightly outperformed the multi-class multi-label BERT-based classifier, suggesting that separating the characterization of different cognitive levels seemed to be a better choice than building only one model to identify all cognitive levels at one time.

Keywords

Learning Objectives, Bloom's Taxonomy, Classification, Machine Learning, BERT

1. INTRODUCTION

A learning objective is a clear and specific statement defining knowledge and skills that learners are expected to acquire after completing an educational activity [19]. A well-articulated learning objective can benefit course designers, instructors and learners. For instance, learning objectives can inform course design, as they often signal how course materials should be organized to ensure a suitable sequencing of instruction and optimize learning activities throughout semester. Instructors can utilize learning objectives to assess learners' progress; meanwhile, learners can use learning objectives to get an overview of knowledge and skills they should possess after receiving instruction [38], and to support their studying for an exam, e.g., by developing questions for self-testing prior to an exam [2].

Educators in many courses create learning objectives that reflect knowledge/skills of different levels of cognitive complexity. For example, evaluating whether a formula from a textbook can be applied to solve a math problem is cognitively a more complex skill compared to recalling that same formula from a textbook. However, a learner needs to be able to recall the formula first, and then evaluate its utility in the context of a genuine problem, i.e., low-order skills are precursors to high-order skills [1, 12]. To define learning objectives at different skill levels, educators often use educational taxonomies (e.g., Bloom's [1, 4], Gagne's [12], and Jensen's [16]). For instance, over decades educators have widely utilized Bloom's taxonomy for Cognitive Objectives [1] to define learning objectives, as this framework can account for a broad range of learning objectives and provide means for evaluating learner achievements relative to those

*Corresponding author.

objectives [17]. Bloom's taxonomy consists of six levels of cognitive skills that include 3 low-order (*remember*, *understand*, and *apply*) and 3 high-order cognitive skills (*analyze*, *evaluate*, and *create*).

Although Bloom's taxonomy has been regarded as a helpful pedagogical framework [19], many educators have difficulties to develop learning objectives appropriate to the levels specified in Bloom's taxonomy [21]. This is due to the fact that they need to consider progression of learners' skills with a learning content and also take into account dependencies between learning objectives, e.g., a learner must be able to define and explain a math formula before applying it [19]. These difficulties may lead to subsequent challenges in measuring learners' progress, i.e., difficulties to determine whether a learner has progressed to upper levels of Bloom's taxonomy [21]. To ensure learning objectives educators create can be mapped to Bloom's taxonomy levels properly, the educators often need support from educational experts [13], which are not easily available in many departments in higher education. Everything considered, the process of developing well-articulated learning objectives to support teaching, learning and assessment activities is usually time- and resource-consuming.

To remedy this challenge and help educators determine the level of each learning objective they create according to Bloom's taxonomy, in the present study, we explored the possibility of using state-of-the-art natural language processing, machine learning and deep learning techniques to automatically classify learning objectives. To date, researchers have developed a few computational models for automatic classification of different types of educational texts based on cognitive levels in Bloom's taxonomy, including exam questions (e.g., [6, 15, 25, 40]), participants' contributions to discussion forums (e.g., [11]), and learning outcomes (e.g, [36]). Researchers have demonstrated a considerable classification accuracy of these models. However, even though classification of learning objectives based on Bloom's taxonomy has been recognized as an important problem, to our knowledge, there has yet to be developed a classification system that accurately automatizes this work. To address this gap, we obtained and manually annotated 21,380 course learning objectives from 5,558 courses from all the 10 constituent faculties at an Australian university, and applied both machine learning and deep learning models to construct classifiers to automatically identify the cognitive levels of these learning objectives.

2. RELATED WORK

2.1 Bloom's Taxonomy

Bloom's taxonomy [1, 4] was originally introduced to reduce educators' labor when preparing the materials for annual comprehensive examinations. The taxonomy proposes six hierarchically-arranged levels of cognition: remember, understand, apply, analyze, evaluate, and create. These levels reflect the cognitive complexity of a learning objective or an assessment question [15]. In particular, remember, understand, and apply are considered low-order, whereas analyze, evaluate, and create are considered high-order cognitive skills. Mastering a skill at a higher level is dependent upon mastering a prerequisite skill or a group of prerequisite skills at lower levels in Bloom's taxonomy. Due to its well-developed structure, educational researchers and practitioners have widely utilized Bloom's taxonomy for both research and instructional purposes, including the classification of learning objectives [36]. We use Bloom's taxonomy as a theoretical framework to guide this study.

2.2 Automated Analysis of Educational Texts Based on Bloom's Taxonomy

Despite its educational promises, the use of Bloom's taxonomy is usually not straightforward. Many educators struggle to manually classify instructional and assessment activities, specify the knowledge associated to each level in Bloom's taxonomy, and measure student progress accordingly ([21]). To overcome these challenges and facilitate instructional activities, researchers have created several computational systems for automated classification of educational texts based on Bloom's taxonomy. Wen-Chih et al. [5] developed a keyword-based system to automatically classify teachers' questions into different cognitive levels of Bloom's taxonomy. For this purpose, the authors developed a dictionary of keywords mapped to the corresponding cognitive levels of Bloom's taxonomy. The classification system developed in this way achieved a considerable accuracy of 75% in identifying questions at the *remember* level, whereas the system's performance in identifying questions at other levels was noticeably lower (25% – 59%). Amali et al. [28] developed several models to automatically classify exam questions into cognitive levels in Bloom's taxonomy. The models included a rule-based part-of-speech classifier, support vector machine, naive Bayes and K-nearest Neighbor classifiers with word vectors as inputs. Similar to [5], the models performed best in identifying exam questions at the *remember* level (87% – 100%) but achieved a lower overall performance (60% – 72%). The authors further created an ensemble model that achieved 82% overall accuracy by combining the four models. Jayakodi et al. [15] utilized semantic similarity algorithms to develop a rule-based classifier that identifies a cognitive level of an exam question according to Bloom's taxonomy. This system achieved a classification accuracy up to 0.70 in identifying a correct cognitive level for an exam question. Similarly, Echeveria et al. [11] computed TF-IDF features in student discussion posts as input for a rule-based classifier that categorizes a post into one of the levels of Bloom's taxonomy. The authors reported the accuracy of nearly 0.77. Waheed et al. [40] and Mohammed et al. [25] developed a group of supervised machine learning models to classify open ended questions according to Bloom's taxonomy. The authors computed a variety of linguistic features from question text, e.g., TF-IDF [32] and word2vec [24]. Whereas these supervised machine learning models were trained using relatively small datasets (i.e., less than 1,000 questions), most of them achieved a substantial classification performance with their F1 score ranging between 0.70 and 0.90.

In line with the increased use of deep learning methods in educational research over the past few years, James et al. [44] utilized BERT [10], a pre-trained language model, to classify educational questions relative to Bloom's taxonomy in a cognitive domain. The models performed well in identifying questions at the levels that were frequent in the dataset (*remember*, *understand*, *analyze* – achieving 82.61% accuracy), whereas the identification of questions at less frequent levels (*apply*, *evaluate*, *create*) remained a challenge (59.2% accuracy with all cognitive levels included). This study demonstrated the potential of deep learning methods to assist educators in determining Bloom's cognitive levels of educational questions if sufficiently big pool of questions at all cognitive levels is available to train the deep learning models. Further, Sarang et al. [36] utilized the pre-trained language model Wiki Word Vectors to generate word embeddings for learning objectives and assessment questions. The word embeddings were used as input to the Long Short Term Memory (LSTM) model classifying learning objectives and assessment questions into different levels of Bloom's taxonomy. This model achieved a weighted average F1 score of 0.73 in correctly classifying learning objectives and a macro-average F1 score of 0.82 in correctly classifying the assessment questions. Moreover, to our knowledge, this study is the first to automatically classify learning objectives (among other forms of educational texts) according to Bloom's taxonomy and the considerable classification performance reported in the study encourages further research. Overall, the classification models developed to date promise to provide at-scale support to educators who aim at categorizing educational texts, e.g., discussion forum posts, assessment questions, and, more recently, learning objectives, based on cognitive levels of Bloom's taxonomy.

Although researchers have begun increasingly harnessing the automated text analysis methods to classify different educational texts based on Bloom's taxonomy, only a small group of researchers has considered exploring the possibility of using these methods to automatically classify learning objectives, despite the challenges documented that many educators report when attempting to manually classify learning objectives according to the cognitive levels of Bloom's taxonomy [21]. Additionally, all of the relevant studies that we found assumed each piece of text to belong to only one cognitive level while the possibility existed for learning objectives to have more cognitive levels as educators could have combined several learning objectives in one sentence. We also note that practical challenges in obtaining a large, manually labeled dataset, with a sufficient number of learning objectives at each of the six levels of Bloom's taxonomy to train the classification models, could have been an important obstacle to this line of research [36]. To address these gaps, we gathered a large amount of authentic learning objectives across different university courses and manually labeled each learning objective with at least one cognitive level in Bloom's taxonomy. Next, we developed classification models based on machine and deep learning methods to automatically classify learning objectives. Specifically, we attempted to answer the following **research question**: *To what extent can machine learning and deep learning classifiers accurately classify a learning objective into the cognitive levels of Bloom's taxonomy?*

3. METHODOLOGY

3.1 Data Collection and Labeling

We collected 21,380 learning objectives publicly available from 5,558 courses provided by the 10 faculties at an Australian university in 2021. To collect the data, we developed a web scraper using Python to automatically parse the content of the available course web pages to obtain learning objectives of a course. One human coder, who had previously received a training on Bloom's taxonomy, manually categorized the learning objectives into their corresponding cognitive levels of Bloom's taxonomy. Some learning objectives were categorized into more than one cognitive level, i.e., 2,325 learning objectives were simultaneously labeled with two cognitive levels, 280 with three and 2 with four levels. We provide a sample of coded learning objectives in Table 1.

To ensure data labeling reliability, a second human coder trained on Bloom's Taxonomy in cognitive domain randomly selected 30% of the learning objectives labeled by the first coder and independently conducted labeling those learning objectives. The two coders achieved a substantial inter-coder agreement (Cohen's κ 0.63), according to the recommendations provided in [22]. The two coders discussed the labeling disagreement cases between them and found out that the major source of disagreement was because many learning objectives from the low-cognition category *remember* were wrongly categorized as the high-cognition category *apply*. The coders revised the corresponding labels in the entire dataset accordingly which increased inter-coder agreement on the 30% of sample to 0.80, measured with Cohen's κ. We proceeded with feature engineering and model development using the labeled dataset to answer our research question. The detailed descriptive statistics is provided in Table 2.

3.2 Classification Models

To answer our research question, we developed and examined six classification models. Of these, five models were based on traditional machine learning algorithms, support vector machine (SVM), logistic regression (LR), naive Bayes (NB), random forest (RF) and XGBoost. These algorithms have been widely utilized for text classification tasks in educational research (for overview see [35]). Moreover, inspired by the increasing use of deep learning approaches in educational research over the past few years (e.g., [7, 9, 14]), we developed deep learning classifiers based on the BERT pre-trained language model. Specifically, we coupled the pre-trained BERT sequence classifiers with a single layer for classification and trained the model using the data we collected.

Recall that, in our collected dataset, each learning objective can be assigned with either only one cognitive label or multiple labels at the same time. Correspondingly, we could tackle the classification task by using two different methods. One is to construct a binary classifier for each cognitive level in Bloom's taxonomy (e.g., those labeled as *remember* vs. those not), and the other is to build one multi-class multi-label classifier for all the cognitive levels in Bloom's taxonomy (i.e., identifying all the cognitive labels specific to a learning objective). As part of our research goal was to shed light on the best way to tackle this problem, we implemented both methods. Specifically, we used each of the models described above to construct a binary classifier for each cognitive level in Bloom's taxonomy, i.e., we constructed a total of 36 binary classifiers. Then, we constructed two multi-class multi-label classifiers based on Random Forest and BERT, as these two models have been demonstrated effective in tackling multi-class multi-label classification problems in previous researches ([27, 37, 39, 42, 43]).

Table 1: Example of learning objectives categorized into the cognitive levels in Bloom's taxonomy.

Learning Objective Examples	Labels
Recognise the key role that human factors play in the leadership and development of a highly functional perioperative team.	Remember
Describe the general characteristics of the modern X-ray system used in clinical practice, including scientific principles, and production of the digital image.	Understand
Apply research skills to operate effectively as a member of a research project team.	Apply
Identify an issue of relevance to the practice of perioperative medicine capable of further investigation and research within the context of a capstone project.	Analyze
Ability to articulate critical interpretations of dramatic texts and processes in systematic written argument.	Evaluate
A capacity to design, manage, and carry out a research project.	Create
Analyze and apply contemporary management theory and research to current organizational issues.	Apply & Analysis
Assess and synthesise diverse information about up-to-date information and knowledge management systems market and how to use implementation strategies to maximise their strengths and minimise their weaknesses.	Evaluate & Create

Table 2: Descriptive statistics of the Learning Objective (LO) dataset.

	Total	Remember	Understand	Apply	Analyze	Evaluate	Create	Multi-Label
# Total LOs	21,380	886	5,079	5,074	2,311	2,468	2,955	2,607
# Avg. words per LO	17.81	16.14	17.42	18.55	16.68	16.64	16.27	21.52
# Avg. unique words per LO	15.75	14.59	15.33	16.40	15.05	14.91	14.74	18.25

3.3 Study Setup

3.3.1 Data Pre-processing

Prior to conducting any experiments, we randomly split the dataset in 80:20 ratio, i.e., 80% of data was used as a training set and 20% data was used as a testing set. We used these same datasets across all classification tasks to ensure fair comparisons between different models.

The textual data was initially pre-processed in the same fashion for both the machine learning models and the BERT-based deep learning model by converting them to lowercase. We extracted multiple features that had been proven to be useful not only for educational forum post classifications [35], but also in other studies sharing similar context to ours ([25], [32]) to empower the five conventional machine learning models described above. In particular, we computed a group of features in n-gram form, including unigrams (1,000 most frequent excluding stopwords) and bigrams (1,000 most frequent excluding stopwords); TF-IDF features (1,000 most frequent excluding stopwords); automated readability index [33] for each learning objective; and 93 features derived from the LIWC dictionary [31] reflecting a frequency of different psychologically meaningful words, e.g., cognitive processes, function words, words reflecting summary, relativity and time orientation, leading to a total of 3,094 features. For our BERT-based deep learning model, unlike some previous studies ([18, 25, 28]) where the researchers used word2vec to generate word embeddings, we employed BERT-uncased shared by HuggingFace [41] to generate word embeddings, because BERT generated embeddings had been proven to be capable of capturing contextual information and properties at the sentence level (in this study, the learning objective level) [23].

3.3.2 Model Implementation

To implement and examine the five conventional machine learning models, we utilized the Python package Scikit-Learn [30] to develop naive Bayes, logistic regression, support vector machine, and random forest classifiers, and the package XGBoost [8] to develop the XGBoost classifier. We performed hyper-parameter tuning with 3-fold cross-validation on these models using grid search in order to find the most suitable parameters for our models. F1 score was used as the evaluation metric when performing hyper-parameter tuning. The details of the parameters were all documented in our source code and would be open-sourced together with the data collected in this study, and thus be made available to other researchers for replications[1]

For the BERT-based model, we applied the BERT-uncased shared by HuggingFace [41]. The model included 12 hidden layers, each with 768 neurons. The vocabulary size was 30,522 and the dropout rate was 0.1. For binary sequence classifiers, the number of output neurons is 2, each predicting the probability of the text belonging to different classes (0 and 1 as class labels). Therefore, we applied a softmax function on these probabilities to find the corresponding class labels for the texts. For the multi-class multi-label sequence classifier, the number of output neurons is 6, predicting the probabilities of the text belonging

[1] https://github.com/SteveLEEEEE/EDM2022CLO.git

to the six cognitive levels. Thus, we used a sigmoid function on these probabilities and set the probability threshold to 0.50 to find their predicted class labels. The entire BERT models were fine-tuned without freezing the parameters in any layers for all experiments.

3.3.3 Model Training

The training data were used to train all the machine learning models without further splitting. However, for our deep learning models, the 80% training data were further split with 80% being training set and 20% being validation set. The batch size was set to 64 for all the deep learning models and the number of epochs was set to 3. Early-stopping was applied in order to avoid over-fitting. When the F1 scores stopped improving on 10 consecutive validations, the training terminated and the model weights rolled back to the best performing one.

3.3.4 Evaluation Metrics

We evaluated the performance of the classification models by computing the following performance metrics: accuracy, Cohen's κ, Area Under the ROC Curve (AUC), and F1 score. To find out the categorical performance on the multi-class multi-label classifiers, we separated the classification results for each of the cognitive levels on testing data and made comparisons with the humanly assigned ones to find out their individual accuracy, Cohen's κ, AUC, and F1 scores.

4. RESULTS

Our results provide evidence that it is possible to develop highly accurate supervised machine learning and deep learning models to classify learning objectives into skill levels based on Bloom's taxonomy, answering our research question. In particular, the high-performing models included those based on SVM, RF, XGBoost and BERT (Table 3). All of the high-performing models achieved Cohen's κ score spanning between 0.79 and 0.93, while the prediction accuracy of these models spanned between 0.92 and 0.99, i.e., the models can accurately classify at least 92% of learning objectives into a corresponding skill level of Bloom's taxonomy. Equally importantly, the F1 scores of the high-performing models were between 0.83 and 0.95, indicating a high precision and recall achieved by SVM, RF, XGBoost, and BERT in identifying each of the six cognitive levels in Bloom's taxonomy. We note that the binary BERT models outperformed all the binary machine learning models observed. These models achieved an outstanding classification performance, as measured by Cohen's κ (0.87 to 0.93), accuracy (0.96 to 0.99), and F1 scores (0.88 to 0.95). The classification performance of other models observed in this study (i.e., naive Bayes and logistic regression) was noticeably lower, e.g., with Cohen's κ typically not exceeding 51% for naive Bayes and 73% for logistic regression.

Furthermore, by comparing the performance between binary and multi-class multi-label random forest classifiers, it is evident that binary random forest classifier outperformed the multi-class multi-label one in most cases except for *understand*. Meanwhile, the multi-class multi-label BERT-based classifier demonstrated to perform better than all the binary and multi-class multi-label machine learning models from the same cognitive level but rarely beat binary BERT-based classifiers in terms of the prediction performance. The exception lied in the cognitive level *evaluate* where binary BERT-based classifier achieved a similar but slightly worse performance than the multi-class multi-label BERT-based classifier (i.e., 0.001 difference in Cohen's κ and F1 score).

5. DISCUSSION

5.1 Interpretation of the Results

Many educators across a range of disciplines develop learning objectives for their courses based on Bloom's taxonomy for Cognitive Objectives [1]. Even though Bloom's taxonomy has been widely deemed a useful pedagogical framework [19], educators often find it challenging and tedious to develop learning objectives to describe cognitive skills at different levels of Bloom's taxonomy [21]. To remedy this issue, in this study, we explored whether machine learning and deep learning methods can be used do develop the classification model that can automatically classify a learning objective into appropriate cognitive level in Bloom's taxonomy. Overall, our results indicated that three traditional machine learning models, i.e., support vector machine, random forest, and XGBoost, and one deep learning model based on BERT, may be the viable approaches towards solving this problem.

The four high-performing classification models achieved considerable performance not only relative to commonly accepted standards in discourse analysis [3], but these models also outperformed the models from prior research that targeted similar classification tasks. Importantly, all the models performed well in correctly classifying learning objectives at each level of Bloom's taxonomy. Given that both conventional machine learning (e.g.,[5, 28]) and deep learning approaches (e.g, [44]) have been documented to perform poorly in classifying educational texts into higher-order cognitive levels of Bloom's taxonomy (e.g., *analyze*, *evaluate*, and *create*), the results of our study add to the body of knowledge in educational research showing that advanced conventional machine learning and deep learning models trained on a large corpus of educational textual data can provide useful classifications across all the levels in Bloom's taxonomy.

Moreover, our findings resonate with prior research showing that deep learning models can provide a more accurate classification results in educational classification tasks, compared to conventional machine learning algorithms [35]. We also note that, given the performance scores the naive Bayes classifier consistently achieved across the six tasks in our study, it appears that this classifier may be the least preferable algorithm for classification tasks based on Bloom's taxonomy, corroborating evidence provided in [29] where the authors pursued the question classification task based on Bloom's taxonomy and found that naive Bayes under-performed other classifiers in this task.

Last but not least, we observed that, though multi-class multi-label classifiers managed to achieve satisfactory performance, binary classifiers using the same model (i.e., BERT) still attained better performance. This might be mainly because that, while multi-class multi-label classifiers tried to minimize the overall errors across different cognitive levels during the model training process, binary classifiers tended to focus comprehensively on minimizing the errors on a single category. Therefore, with adequate data collected, tack-

Table 3: Classification Performance of the binary and multi-class multi-label (MCML) classifiers, i.e., Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), XGBoost, and the BERT-based classifier. The best results are in bold for each evaluation metric in each level of Bloom's taxonomy.

Methods		Remember				Understand				Apply			
		Acc.	Cohen's κ	AUC	F1	Acc.	Cohen's κ	AUC	F1	Acc.	Cohen's κ	AUC	F1
Binary Classifiers	NB	0.640	0.111	0.716	0.198	0.642	0.327	0.716	0.581	0.778	0.507	0.781	0.668
	SVM	0.982	0.827	0.923	0.837	0.922	0.801	0.891	0.855	0.923	0.805	0.890	0.858
	LR	0.960	0.485	0.681	0.503	0.891	0.714	0.839	0.787	0.896	0.726	0.837	0.793
	RF	0.983	0.830	0.892	0.839	0.920	0.793	0.880	0.847	0.936	0.837	0.904	0.881
	XGBoost	0.981	0.820	0.916	0.830	0.928	0.818	0.900	0.867	0.938	0.844	0.914	0.887
	BERT	**0.987**	**0.871**	0.916	**0.878**	**0.971**	**0.926**	0.959	**0.947**	**0.961**	**0.904**	0.951	**0.931**
MCML Classifiers	RF	0.982	0.809	**0.989**	0.818	0.927	0.811	0.970	0.860	0.921	0.794	0.970	0.847
	BERT	0.984	0.848	0.988	0.856	0.955	0.889	**0.982**	0.920	0.951	0.877	**0.976**	0.912

Methods		Analyze				Evaluate				Create			
		Acc.	Cohen's κ	AUC	F1	Acc.	Cohen's κ	AUC	F1	Acc.	Cohen's κ	AUC	F1
Binary Classifiers	NB	0.549	0.183	0.684	0.392	0.596	0.234	0.703	0.447	0.676	0.300	0.743	0.474
	SVM	0.956	0.832	0.897	0.858	0.959	0.861	0.922	0.886	0.942	0.791	0.880	0.825
	LR	0.936	0.732	0.818	0.767	0.920	0.694	0.799	0.739	0.902	0.604	0.762	0.659
	RF	0.961	0.851	0.902	0.874	0.967	0.887	0.932	0.907	0.943	0.792	0.877	0.826
	XGBoost	0.959	0.844	0.903	0.868	0.964	0.878	0.924	0.900	0.944	0.796	0.882	0.829
	BERT	**0.975**	**0.906**	0.950	**0.922**	0.974	0.913	0.954	0.929	**0.962**	**0.866**	0.924	**0.888**
MCML Classifiers	RF	0.951	0.803	0.972	0.831	0.950	0.822	0.981	0.852	0.928	0.715	0.964	0.755
	BERT	0.971	0.890	**0.984**	0.907	**0.974**	**0.914**	**0.989**	**0.930**	0.958	0.846	**0.971**	0.872

ling the problem as multiple binary classification tasks may be a better solution.

5.2 Practical Implications

In this study, we made a first step towards developing future computational tool that can provide at-scale support to instructors, instructional designers, and other educational stakeholders who aim at developing learning objectives well aligned to Bloom's taxonomy. The system will automatically analyze learning objectives using the classification routines developed in this study. For instance, an instructional designer may submit the list of manually created course learning objectives to this future system and obtain a highly accurate classification of the learning objectives into cognitive levels of Bloom's taxonomy. Using this information, the instructional designer may determine whether all the learning objectives are provided, relative to course requirements, e.g., "*It looks like I yet to develop a learning objective at the create level. Since this is an advanced writing course, the create learning objectives should be included*" or "*Even though I have created a few learning objectives for the skills at the apply level, my list is missing lower-level learning objectives that represent the corresponding pre-requisite skills*". Overall, the classifiers developed in this study can be used to automatically diagnose the cognitive levels of learning objectives for courses and educational programs across different faculties as well as universities.

In addition, coupled with the systems for natural language generation, the classifiers of learning objectives might be further enhanced to automatically generate learning objectives from course content. This, in turn, may reduce time educators dedicate to this task and may mitigate inconsistencies educators introduce among each other when defining learning objectives, e.g., two instructors defining different learning objectives for the same subject. We also anticipate our work will benefit students by providing means for automatic development of questions of different cognitive levels for self-assessment. For example, automatically generated learning objectives can be further coupled with the systems for automatic question generation to obtain interrogative form for objectives. Questions developed in this way may provide at-scale support to students studying for assessment purposes, e.g., those in Massive Open Online Courses.

6. LIMITATIONS AND FUTURE WORK

We identified several limitations in this study that may be considered in future research. Firstly, even though all the learning objectives we collected were classified into at least one of the cognitive levels in Bloom's taxonomy, it is, however, possible that some learning objectives cannot be categorized relative to a cognitive domain but relative to other domains instead, e.g., affective domain [26]. In future research, the learning objectives dataset should be further labeled from other domains, and relevant classifiers should be trained to recognize these types of learning objectives. Secondly, the supervised machine learning and deep learning methods utilized in this study require extensive amounts of labeled data to achieve a highly accurate prediction performance. As preparing such a large-scale dataset can be costly and time-consuming, researchers may consider using semi-supervised machine learning approaches (e.g., semi-supervised Random Forest [20]) or training strategies like active learning [34] to enable more effective and efficient model construction process in the future.

7. REFERENCES

[1] L. W. Anderson and D. R. Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman,,

2001.

[2] H. L. Andrade. A critical review of research on student self-assessment. In *Frontiers in Education*, volume 4, page 87. Frontiers, 2019.

[3] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596, 2008.

[4] B. S. Bloom et al. Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, 20(24):1, 1956.

[5] W.-C. Chang and M.-S. Chung. Automatic applying bloom's taxonomy to classify and analysis the cognition level of english question items. In *2009 Joint Conferences on Pervasive Computing (JCPC)*, pages 727–734, 2009.

[6] G. Chen, J. Yang, C. Hauff, and G.-J. Houben. Learningq: a large-scale dataset for educational question generation. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[7] J. Chen, J. Feng, X. Sun, and Y. Liu. Co-training semi-supervised deep learning for sentiment classification of mooc forum posts. *Symmetry*, 12(1):8, 2020.

[8] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[9] B. Clavié and K. Gal. Edubert: Pretrained deep language models for learning analytics. *arXiv preprint arXiv:1912.00690*, 2019.

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

[11] V. Echeverría, J. C. Gomez, and M.-F. Moens. Automatic labeling of forums using bloom's taxonomy. In *International Conference on Advanced Data Mining and Applications*, pages 517–528. Springer, 2013.

[12] R. M. Gagné, R. M. Gagné, et al. *Conditions of learning and theory of instruction*. Holt, Rinehart and Winston, 1985.

[13] R. Gluga, J. Kay, R. Lister, Simon, and S. Kleitman. Mastering cognitive development theory in computer science education. *Computer Science Education*, 23(1):24–57, 2013.

[14] S. X. Guo, X. Sun, S. X. Wang, Y. Gao, and J. Feng. Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in mooc discussion forums. *IEEE Access*, 7:120522–120532, 2019.

[15] K. Jayakodi, M. Bandara, I. Perera, and D. Meedeniya. Wordnet and cosine similarity based classifier of exam questions using bloom's taxonomy. *International Journal of Emerging Technologies in Learning*, 11(4), 2016.

[16] A. Jensen. Varieties of individual differences in learning. in, rm gagne. *Learnin˜ and Individual Differences. Columbus, Ohio: Merrill Books*, 1967.

[17] P. C. Kyllonen and V. J. Shute. Taxonomy of learning skills. Technical report, UNIVERSAL ENERGY SYSTEMS INC DAYTON OH, 1988.

[18] M. Laddha, V. Lokare, A. Kiwelekar, and L. Netak. Classifications of the summative assessment for revised bloom's taxonomy by using deep learning. *International Journal of Engineering Trends and Technology*, 69:211–218, 03 2021.

[19] M. B. Larson and B. B. Lockee. *Streamlined ID: A practical guide to instructional design*. Routledge, 2019.

[20] C. Leistner, A. Saffari, J. Santner, and H. Bischof. Semi-supervised random forests. In *2009 IEEE 12th international conference on computer vision*, pages 506–513. IEEE, 2009.

[21] S. Masapanta-Carrión and J. Á. Velázquez-Iturbide. A systematic review of the use of bloom's taxonomy in computer science education. In *Proceedings of the 49th acm technical symposium on computer science education*, pages 441–446, 2018.

[22] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[23] A. Miaschi and F. Dell'Orletta. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online, July 2020. Association for Computational Linguistics.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[25] M. Mohammed and N. Omar. Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PloS one*, 15(3):e0230442, 2020.

[26] R. W. Morshead. Taxonomy of educational objectives handbook ii: Affective domain. *Studies in Philosophy and Education*, 4(1):164–170, 1965.

[27] V. Nikolovski, D. Kitanovski, D. Trajanov, and I. Chorbev. Case study: Predicting students objectivity in self-evaluation responses using bert single-label and multi-label fine-tuned deep-learning models. In V. Dimitrova and I. Dimitrovski, editors, *ICT Innovations 2020. Machine Learning and Applications*, pages 98–110, Cham, 2020. Springer International Publishing.

[28] A. Osadi, N. Fernando, and V. Welgama. Ensemble classifier based approach for classification of examination questions into bloom's taxonomy cognitive levels. *International Journal of Computer Applications*, 162:975–8887, 04 2017.

[29] A. Osman and A. A. Yahya. Classifications of exam questions using natural language syntatic features: A case study based on bloom's taxonomy. In *Proc. 6th Int. Arab Conf. Quality Assurance Higher Edu.*, 2016.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[32] C. Sammut and G. I. Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[33] R. Senter and E. A. Smith. Automated readability index. Technical report, Cincinnati Univ OH, 1967.

[34] B. Settles. Active learning literature survey. 2009.

[35] L. Sha, M. Rakovic, Y. Li, A. Whitelock-Wainwright, D. Carroll, D. Gaševic, and G. Chen. Which hammer should i use? a systematic evaluation of approaches for classifying educational forum posts. *International Educational Data Mining Society*, 2021.

[36] S. Shaikh, S. M. Daudpotta, and A. S. Imran. Bloom's learning outcomes' automatic classification using lstm and pretrained word embeddings. *IEEE Access*, 9:117887–117909, 2021.

[37] S. Sharma and D. Mehrotra. Comparative analysis of multi-label classification algorithms. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pages 35–38, 2018.

[38] H. Sullivan and N. Higgins. *Teaching for competence.* ERIC, 1983.

[39] T. Tang, X. Tang, and T. Yuan. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8:193248–193256, 2020.

[40] A. Waheed, M. Goyal, N. Mittal, D. Gupta, A. Khanna, and M. Sharma. Bloomnet: A robust transformer based model for bloom's learning outcome classification. *arXiv preprint arXiv:2108.07249*, 2021.

[41] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[42] R. Yarullin and P. Serdyukov. Bert for sequence-to-sequence multi-label text classification. In W. M. P. van der Aalst, V. Batagelj, D. I. Ignatov, M. Khachay, O. Koltsova, A. Kutuzov, S. O. Kuznetsov, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, M. Pelillo, A. V. Savchenko, and E. Tutubalina, editors, *Analysis of Images, Social Networks and Texts*, pages 187–198, Cham, 2021. Springer International Publishing.

[43] H. M. Zahera. Fine-tuned bert model for multi-label tweets classification. In *TREC*, 2019.

[44] J. Zhang, C. Wong, N. Giacaman, and A. Luxton-Reilly. Automated classification of computing education questions using bloom's taxonomy. In *Australasian Computing Education Conference*, ACE '21, page 58–65, New York, NY, USA, 2021. Association for Computing Machinery.