Toward the Automatic Labeling of Course Questions for Ensuring their Alignment with Learning Outcomes

S. Supraja
Nanyang Technological University
50 Nanyang Ave
Singapore 639798
ssupraja001@e.ntu.edu.sg

Kevin Hartman
Nanyang Technological University
50 Nanyang Ave
Singapore 639798
khartman@ntu.edu.sg

Sivanagaraja Tatinati
Nanyang Technological University
50 Nanyang Ave
Singapore 639798
tatinati@ntu.edu.sg

Andy W. H. Khong
Nanyang Technological University
50 Nanyang Ave
Singapore 639798
andykhong@ntu.edu.sg

ABSTRACT

Expertise in a domain of knowledge is characterized by a greater fluency for solving problems within that domain and a greater facility for transferring the structure of that knowledge to other domains. Deliberate practice and the feedback that takes place during practice activities serve as gateways for developing domain expertise. However, there is a difficulty in consistently aligning feedback about a learner's practice performance with the intended learning outcomes of those activities – especially in situations where the person providing feedback is unfamiliar with the intention of those activities. To address this problem, we propose an intelligent model to automatically label opportunities for practice (assessment questions) according to the learning outcomes intended by the course designers. As a proof of concept, we used a reduced version of Bloom's Taxonomy to define the intended learning outcomes. Using a factorial design, we employed term frequency-inverse document frequency (TF-IDF) and latent Dirichlet allocation (LDA) to transform questions from text to word weightages with support vector machine (SVM) and extreme learning machine (ELM) to train and automatically label the questions. We trained our models with 120 questions labeled by the subject matter expert of an undergraduate engineering course. Compared to existing works which create models based on a self-generated dataset, our proposed approach uses 30 untrained questions from online/textbook sources to validate the performance of our models. Exhaustive comparison analysis of the testing set showed that TF-IDF with ELM outperformed the other combinations by yielding 0.86 reliability (F1 measure) with the subject matter expert.

Keywords

Learning outcomes, Term frequency-inverse document frequency, Latent Dirichlet allocation, Extreme learning machine, Support vector machine

1. INTRODUCTION

Increasingly, modern curriculum design in tertiary and adult learning settings has become a collaborative endeavor between subject matter experts, learning designers, and learning technologists. While these teams employ a variety of process models for the planning, execution, and revision of their curriculum and activity designs, often greater attention is paid to the construction of a course design and the course content rather than the assessment practices that measure learning and their ongoing maintenance.

The algorithms and use case described in this paper exist in a particular context of outcome-based education. In this context, learning is defined by observable changes in a learner's behavior. These changes commensurate with Krathwohl's model of learning objectives [1] but learning outcomes go beyond objectives. Learning outcomes are predicated on having learners observably demonstrate their growing understanding of a topic or proficiency within a field [2]. When learning activities become more open-ended and exploratory, and when learners are offered choices for how to proceed, learners often look to how they will ultimately be assessed to gauge which learning strategies they should employ [3].

When a course's learning activities support its assessment practices and the assessment practices support the types of outcomes that are relevant to learners in the future, the course's activities and intended learning outcomes exhibit constructive alignment with each other [2]. Adhering to constructive alignment creates a seamless path from learning, to applying, to transferring concepts and relationships when solving novel problems.

However, the promise of constructive alignment is not easily delivered upon. Oftentimes, a course's learning outcomes cannot be measured by its assessment practices, or its assessment practices are decontextualized from the types of activities and practices learners are actually preparing for [4]. Whether in the context of higher learning or professional development, when thinking about developing flexible, life-long learners it is paramount to have mechanisms in place to support learners as they work to gain domain expertise. These processes should reliably measure learning and link assessment practices to authentic activities.

1.1 Learning design for domain expertise

Prior work in designing for adaptive domain expertise, the kind of expertise necessary for learners to function in changing environments and flexible job scopes, has shown that learning design teams need to be cognizant of three elements which will be discussed in turn.

1.1.1 Levels of learning outcomes

Learning outcomes range in sophistication and vary by field. In medicine, Miller's Pyramid [5] lists learning outcomes beginning with knowing about a subject, progressing to knowing how to do something, to being able to actually demonstrate it in a contrived setting like a role-play with actors, and to being able to demonstrate it in a real environment like a surgical theater [6]. The idea is based on the belief that the development of expertise is a progression from

the recall of facts to the execution of skills. However, as research on problem based learning has shown, demonstration of skill and the recall of facts can proceed independently of each other depending on the learning environment [7].

In [8], a field agnostic method of classifying learning outcomes based on their quality is presented. Essentially, the Structure of Observed Learning Outcomes (SOLO) taxonomy identifies the level of cognitive sophistication a learning outcome requires. Lower level learning outcomes indicate a learner is capable of remembering facts in isolation. More sophisticated levels require learners to assimilate information from various sources to make connections and transform that understanding into something new.

Perhaps the most popular listing of learning outcomes is Bloom's Taxonomy. Similar to Miller's Pyramid, Bloom's Revised Taxonomy also begins with the retrieval of facts and information as its foundation and builds up to application of knowledge and further to analyzing, evaluating, and creating. Because of its simplicity and familiarity with learning designers and subject matter experts alike, Bloom's Taxonomy can easily be used to identify the levels of learning outcomes in a course [9].

1.1.2 Opportunities for deliberate practice

Along with identifying a learning activity's intended outcomes, expertise development requires opportunities for deliberate practice. In contrast to repetitive practice intended for learners to develop automaticity in either the recall of information or the application of a skill, often during time-limited tasks, deliberate practice focuses on mastering the nuances of the domain itself to fine-tune performance [10]. In fact, a learner's level of grit, a combination of perseverance and passion, predicts how close to expert performance a learner will eventually show [11].

The key difference in processes between repetitive practice and deliberate practice leads to different forms of expertise: adaptive and routine [12]. Routine forms of expertise allow a learner to conduct a task at an optimal level. Adaptive expertise allows learners to learn new tasks or solve novel problems at an accelerated rate. In an industrial setting, routine expertise helps a worker complete a particular job function. Adaptive expertise enables that same worker to retrain to fill new job functions. Typically, the amount of time necessary to achieve expert performance in a domain is in the order of years to decades [13]. However, incremental improvement can be seen in a few practice cycles when activities align to the intended learning outcomes.

1.1.3 Formative assessments and actionable feedback

Hand in hand with creating opportunities for deliberate practice is providing formative feedback to the learner about how to improve that practice while that improvement is still relevant. Imagine students who diligently answer every question in an engineering textbook but never receive feedback on the quality of their solutions. In this case, the learners would be unable to gauge their performance in relation to the course learning outcomes or have an idea about how to improve their performance in the future. Now imagine if those same students do receive feedback, but that feedback arrives after the course's final examination. If the content of the course is mostly self-contained and will not be revisited, the feedback is mostly irrelevant.

Formative feedback consists of two parts: 1) an interpretable indication of a learner's performance on an assessment of learning with respect to a standard of performance (learning outcome) and 2) the opportunity to improve performance before the final evaluation [14].

Cognitive tutors provide a clear example of the power of coupling formative assessment and actionable feedback together in the domain of mathematics learning [15]. By presenting learners with a series of structured problems, cognitive tutors are capable of intervening at any point during the problem-solving process to provide students with feedback about their performance. This feedback may be the identification of an error, the presentation of a hint, or the request for more information about the learner's reasoning. After the feedback, learners have the opportunity to adjust their problem-solving heuristics to improve their performance going forward.

Such an interaction sequence works with highly structured tasks with application-oriented learning outcomes. However, the feedback cycle is more difficult to manage when the learning outcomes are aligned to higher-order reasoning like evaluation, analyzing and creating. These outcomes have multiple paths for reaching a satisfactory answer.

With this difficulty in mind, we looked at techniques to automate the process of identifying the reasoning level of text-based assessment items (questions) with the intention of better aligning questions to learning outcomes as a first step toward being able to provide opportunities for deliberate practice. Subsequently, the outcome of our proposed work is to link actionable feedback to a learner's performance on assessment items.

1.2 Automated question classification techniques

Prior work has shown the viability of automatically labeling questions in accordance with a course's learning outcomes. However, our work goes beyond labeling existing content to helping course instructors promote deliberate practice and expertise development by providing a method of finding new questions that align to the course designer's original intended learning outcomes. We highlight the drawbacks of prior work and how our proposed approach addresses those limitations.

1.2.1 Labeling questions based on difficulty level

Early attempts at automatically labeling questions relied on subject matter experts to pre-define the difficulty levels of questions. Artificial neural network trained by backpropagation then used the question features and assigned difficulty levels in the training set to classify new questions. A five-dimensional feature vector that consisted of query-text relevance, mean term frequency, length of questions and answers, term frequency distribution (variance), distribution of questions and answers in a text were used. The method yielded an F1 measure, a classification reliability metric that measures a test's accuracy, of 0.78 [16]. However, a major pitfall this method is its lack of semantic analysis.

Entropy-Based Decision Tree has also been used to label questions [17]. The weakness in this strategy is that there is high possibility of overfitting the model during the training phase that then negatively affects the subsequent prediction performance.

1.2.2 Labeling questions based on Bloom's Taxonomy using Natural Language Processing

Natural Language Processing (NLP) has been used for the generation of assessments, answering questions, supporting users in Learning Management Systems and preparing course materials. The Wordnet package has been used to detect semantic similarity. By performing a rule-based approach, the accuracy of labeling a

question based on Bloom's Taxonomy reaches 82% [18]. To improve the rule-based approach, a hybrid technique of using an N-gram classifier with a rule-based approach has also been explored. Rules were based on combining parts-of-speech tagging, and the N-gram classifier found the probabilities of predicting certain words. Such a hybrid method yielded an F1 measure of 0.86 [19].

1.2.3 Labeling questions based on Bloom's Taxonomy using machine learning techniques

Machine learning algorithms can be broadly split into either supervised or unsupervised training implementations. Generally, supervised training is adopted when, during training, labels have been pre-determined and questions are labeled by an expert. The most commonly used method in such cases is the term frequency-inverse document frequency (TF-IDF). The algorithm assigns weightages to individual words in a question statement to define a custom vector space to each question.

Machine learning techniques such k-nearest neighbors, Naïve Bayes and support vector machine (SVM) have been implemented for labeling questions. When doing a performance comparison among these three techniques, an F1 measure of 0.71 was achieved using SVM [20]. To increase the accuracy level, additional features were incorporated in future versions of the work. Three different feature selection processes, namely: Odd Ratio, Chi-square statistic and Mutual Information were used with the three machine learning techniques. The F1 measure result reached 0.9 [21].

Furthermore, an integrated approach of feature extraction has been proposed by using headword, semantic, keyword and syntactic extractions, which are fed into SVM [22]. However, this work has not yet been completed by using a testing dataset to quantify the reliability of prediction.

A major downside in existing works is that both the training as well as testing questions are part of the same course curriculum; the questions are generated by the same author/instructor. Even when a high F1 measure is achieved, it does not enable the algorithm to label questions written by another subject matter expert. Our work increases the flexibility of labeling methods by testing our models with a new set of questions compiled from textbook and online resources.

In addition, our work introduces extreme learning machine (ELM), which has been shown to outperform SVM during similar labeling tasks [23]. Moreover, we introduce LDA as an alternative technique to TF-IDF for transforming question statements into numerical word weightages.

By comparing combinations of these new techniques with more traditional techniques, we aim to gauge which combination attains the highest labeling reliability with the subject matter expert when automatically labeling untrained questions. For our purposes, using the combination with the highest F1 measure (fewest false negatives and false positives) becomes paramount. In our use case, a mislabeling by the algorithm will lead to the wrong set of practice questions to be given to students and diminish the impact of deliberate practice on reaching the intended learning outcomes.

2. METHODS

2.1 Materials

2.1.1 Labeling scheme

The core of this study centers on a labeling scheme for identifying the sophistication of learning outcomes based on a simplified version of Bloom's Taxonomy. In this labeling scheme, the first two levels of Bloom's Taxonomy (Remembering and Understanding) were collapsed into Remember. Applying remained its own category. All of the higher-order reasoning categories (Analyzing, Evaluating, and Creating) were collapsed into Transfer. Figure 1 shows how our labeling scheme categories map onto the original categories from Bloom's Revised Taxonomy.

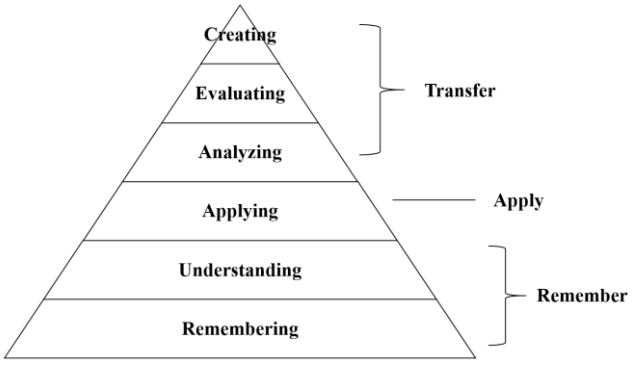

Figure 1: Mapping of Bloom's Revised Taxonomy [24]

We collapsed the taxonomy into three categories for two reasons. First, the subject matter expert tasked with labeling the questions was unsure about how reliably the questions could be labeled by someone without a background in learning design, educational psychology, or curriculum development. Collapsing the categories to Remember, Apply, and Transfer made manually labeling hundreds of questions to train the machine learning algorithms more tractable. Second, collapsing the categories had the effect of making Bloom's Taxonomy more analogous to the successful use cases of Miller's Pyramid by subject matter experts in both higher education and professional development settings [5].

2.1.2 Question dataset

The dataset consists of a total of 150 questions used for training and testing the machine learning algorithms based on the content of an undergraduate electrical and electronic engineering course.

For this study, we formed a training set of 120 questions by randomly selecting 40 Remember, Apply, and Transfer items from the larger question pool of more than 200 questions used in that course. The pool came from a repository of four years' worth of assignment, homework, quiz and exam questions presented to students. These questions prompt students for a range of answer types (i.e., open-ended, multiple-choice, short-structured, essay).

We then created a testing set of 30 new questions compiled from external sources such as textbooks and online question banks. This set was also balanced with equal representation of Remember, Apply, and Transfer questions.

2.2 Data pre-processing procedures

We pre-processed the raw questions in two phases. First, the subject matter expert labeled every question according to the labeling scheme described above. Second, we transformed the text of every question into a machine-readable format before passing them through the machine learning algorithms.

2.2.1 Subject matter expert pre-processing

The subject matter expert manually labeled each question in the training set based on its intended learning outcome (Remember, Apply or Transfer). The subject matter expert then labeled the 30 new questions in the testing set in the same manner. These new questions are labeled for the purpose of knowing the ground truth for performance evaluation. Table 1 below shows some examples of the labeled questions.

Table 1 - Examples of labeled questions

Remember
Consider a signal described by y[n] = 2n +4. What would be the amplitude of the signal at sample index n=3?
Apply
Consider the following input and output signals: find the transfer function and state the poles and zeros of this transfer function.
Transfer
Describe how the bandpass filter can be utilized for radar applications.

2.2.2 Text pre-processing

The text transformation began by excising all equations, mathematical symbols and diagrams from the questions. We only kept the core of the question prompts by removing the descriptive and explanatory text from scenario and hypothetical questions. For example, if a question began by setting the stage with "Peter has been asked to perform…", followed by the question prompt "How much voltage should Peter expect in the circuit?", all of the descriptive text prior to the question prompt was removed to improve the consistency of word length and usage between items.

For the remaining words in the questions, we changed all of the characters to lower case, removed all punctuation marks, numbers, and non-unicode characters. We then stemmed the remaining words to obtain a list of root words. From this list of root words, we removed all words with fewer than three letters. Because we were unsure of the relationship between the words and the labels, we did not create a list of stopwords for removal.

3. TECHNIQUES

We tested four combinations (in no particular order) of word weighting and question labeling algorithms, as shown in Figure 2, to identify the techniques with the highest reliability for our automated learning outcome labeler.

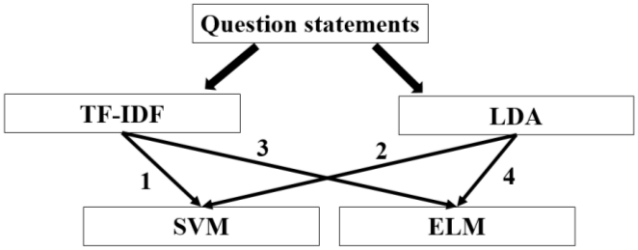

Figure 2: Four combinations of algorithms

Every word in each question prompt was assigned a weightage value based on either term frequency-inverse document frequency (TF-IDF) or latent Dirichlet allocation (LDA). Subsequently, the vector values for each question were passed through either support vector machine (SVM) or extreme learning machine (ELM) to assign a label. All algorithms were implemented in R Studio.

3.1 Term frequency-inverse document frequency

Term frequency-inverse document frequency (TF-IDF) is a technique for finding the relative frequency of words in a given document, and comparing those frequencies with the inverse of how often each of those words appear in the complete document corpus. The resulting ratio can be used to signify the relevance of each unique word within a single document.

We implemented a modified version of TF-IDF that used individual questions as the source of the analysis instead of complete documents. This focused the model on finding the relevance of each word within each single question. By converting each question into a vector of weightages based on word frequencies, the machine learning algorithms were then used to label the questions. The modified TF-IDF model can be described by

$$TF - IDF(w_i, q_k) = \#(w_i, q_k) \times \log \frac{TR}{\#TR(w_i)} \quad (1)$$

where w_i refers to a particular word i, q_k refers to a particular question k, $\#(w_i,q_k)$ refers to number of times w_i occurs in q_k, TR refers to total number of questions and #TR(w_i) refers to question frequency, or the number of questions in which w_i occurs [20].

In the case where the term frequency (TF) count is biased towards longer questions, the TF count is normalized as

$$TF_{i,k} = \frac{n_{i,k}}{\sum_j n_{j,k}} \quad (2)$$

where $n_{i,k}$ refers to the number of times w_i occurs in q_k, the denominator term (size of each question) refers to the sum of the number of times each word appears in q_k [25].

For our work, the pre-processing procedures registered a total of 465 unique stemmed words in our compilation of 120 training questions and 30 testing questions. This led to each question being represented as a vector of 1 row and 465 columns arranged in alphabetical order by stemmed word. When a word is present in a question, the normalized weight of that word is assigned to that question's vector element. If a word is not present in the question, the weight is zero.

After determining the unique word weightage vectors for all 150 questions, the entire matrix is sorted such that for each question, the weightages are arranged in ascending order. The top ten weightages are chosen for each question. The 10 weightages may correspond to different words in each question, but their combinations remain question-specific and give a numerical representation of each question statement. This new vector of 10 columns per question serves as the input to the machine learning algorithms.

As an example, we will use the pre-processed question prompt:

for signal which begin when the one side unilateral ztransform given

Table 2 below shows the weightages assigned to the above example after the application of the TF-IDF technique. The weightages are then arranged in ascending order and the top 10 values are taken.

Table 2 - TF-IDF weightage arrangement

Word (alphabetical order)	**Weightage**
begin	0.392
for	0.140
given	0.140
one	0.222
side	0.356
signal	0.116
the	0.007
unilateral	0.392
when	0.279
which	0.230
ztransform	0.216

3.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a probabilistic technique for topic modeling based on the Bayesian model. The essential idea of LDA is that each document consists of a mixture of topics, with the continuous-valued mixture properties distributed in a Dirichlet random variable, a continuous multivariate probability distribution.

Again, in the context of our work, we applied LDA to questions in the dataset by substituting the original notion of documents in the LDA algorithm with questions in our modified model. Therefore, the modified model attempted to find *k* number of topics (*k* is a user-defined parameter to determine the desired number of topics, or dimensionality of the Dirichlet distribution) for a given set of question statements based on the choice and usage of words in each question. The joint distribution of a topic mixture, a set of topics and a set of words can be represented by

$$p(\theta, t, w|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{M} p(t_i|\theta) p(w_i|t_i, \beta) \quad (3)$$

where parameter α is a *k*-vector with components more than zero, parameter β refers to the matrix of word probabilities, θ refers to a *k*-dimensional Dirichlet random variable, t_i refers to a topic, w_i refers to a word [26].

Figure 3 shows a graphical model representation of LDA. The bigger circle refers to questions while the smaller circle refers to the repeated choice of topics and words within each question.

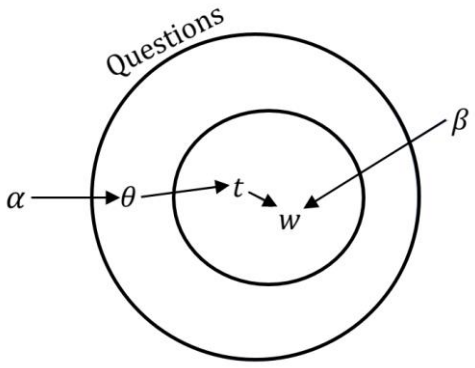

Figure 3: Graphical model representation of LDA

Since LDA involves topic modeling, an appropriate *k* value chosen for our work was ten. This allowed a standard comparison between LDA and the top ten weightages from the TF-IDF method. The generated unique topics (based on the stemmed words) are shown in Table 3.

Table 3 - Topic names generated by LDA

Topic number	Stemmed topic name
1	differ
2	discrete
3	impulse
4	signal
5	filter
6	apply
7	dft
8	output
9	sample
10	system

Out of the entire set of stemmed words detected, ten words have been identified as topic names. Hence, LDA automatically associates the remaining words the above-mentioned ten topics. Based on the words that appear in each question, LDA displays the number of topics per question. Based on the topic assignments, the topic weightages for each question is generated. For topics not present in a question, a minimal weightage is given to those topics in lieu of a zero value. The value ensures that the topic weightages for a question sum to one. Similar to the TF-IDF output, the new vector of 10 columns per question becomes the input for the machine learning algorithms.

3.3 Extreme learning machine

Extreme learning machine (ELM) is a learning algorithm for single-hidden layer feedforward neural networks (SLFNs). ELM can be used for classification, regression, clustering, compression and feature learning. ELM randomly chooses the hidden nodes and determines the output weights of the neural networks.

The following three-step learning model explains ELM. Given a training set that is labeled (information about the target nodes), hidden node activation function and number of hidden nodes,

Step 1: Randomly assign hidden node parameters

Step 2: Calculate the hidden layer output matrix, **H**

Step 3: Calculate the output weight γ

Given a set of inputs with unknown labels, the objective is to find the target outputs [27]. Once the inter-layer weights have been found, the same weights are used during the testing phase. For a given set of input samples x_k, the target/output is given by t_k. For number of hidden nodes *L* and with a certain activation function *f(x)*, the SLFN is modeled as

$$\sum_{j=1}^{L} \gamma_j f_j(x_k) = \sum_{j=1}^{L} \gamma_j f\left(w_j \cdot x_k + b_j\right) = o_k, k = 1, \dots, L \quad (4)$$

where w_j refers to the weight vector that stores the weights between input and hidden nodes, γ_j refers to the weight vector that stores the weights between the hidden and output nodes, b_j refers to the threshold of the jth hidden nodes. The objective is that o_k and t_k (original target) should have zero difference [23] using possible activation functions that include sigmoid, sine, radial basis and hard-limit.

In our case, the output of the ELM are three continuous values that represent the values assigned to the three learning outcome categories (Remember, Apply and Transfer). To convert the three values into a binary value for comparing the predicted labels with the actual labels, we set the learning outcome category with the highest value to one and the remaining two to zero.

3.4 Support vector machine

Support vector machine (SVM) is a mapping of data samples such that these samples can be distinctly labeled. The concept of SVM is derived from margins and subsequently separating data into groups with large gaps between them. Deriving an optimal hyperplane for identifying linearly separable patterns is the key to SVM. This idea is extended to cases where the patterns are non-linearly separable, by using a kernel function to transform the original data samples to map onto a new space [28]. Possible kernels are: linear, polynomial, radial basis and sigmoid.

For our work, we used the C-support vector classification type. Given a set of inputs and targets, the cost function is given by [29]

$$\min_{p,m,\xi} \frac{1}{2} p^T p + C \sum_{j=1}^{k} \xi_j \quad (5)$$

subject to $y_j\left(p^T \phi\left(v_j\right) + m\right) \geq 1 - \xi_j, \xi_j \geq 0, j = 1, \dots, k$

where C>0 is the regularization parameter, m is a constant, p is the vector of coefficients, ξ_j refers to parameters that handle the inputs, index j refers to labeling the k training cases, v refers to the independent variables, y refers to the class labels, ϕ refers to the kernel used that transforms data from the input to the chosen feature space.

Fundamentally, support vectors are data points that lie close to the decision boundary, which are the hardest to classify. SVM maximizes the margin around the hyperplane that separates these points. The cost function is determined based on the training samples (support vectors). These support vectors are the basic elements of a training set that would change the position of the hyperplane dividing the dataset. SVM becomes an optimization problem for determining the optimal hyperplane.

3.5 Performance metrics

To evaluate the reliability of our four technique combinations with the subject matter expert's labels, we looked at using the F1 measure. Accuracy is the number of correct labels divided by the size of testing data. The F1 measure is a harmonic mean of two other metrics: precision and recall. Precision refers to the correctness of questions that have been selected as a particular category. Recall refers to the correctness of selection of the correct category given all the questions that were correctly classified.

Because minimizing the number of false positives and false negatives was important for accurately assigning new questions to the correct practice sets, we used the F1 measure as the basis for our algorithm comparisons. To explain the F1 measure, we will step through the confusion matrix used to describe the performance of a labeling model on a set of testing data. There are four concepts used to construct the confusion matrix:

True positive (TP) refers to the number of questions that the algorithm correctly identifies as presenting a label.

False positive (FP) refers to the number of questions that the algorithm identifies as presenting a label while the subject matter expert indicates the label was absent.

True negative (TN) refers to the number of questions that the algorithm correctly identifies as having a label absent.

False negative (FN) refers to the number of questions that the algorithm identifies as having a label absent while the subject matter expert indicates the label was present.

The F1 measure is calculated as follows [30]

$$Precision = \frac{TP}{(TP+FP)} \quad (6)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (7)$$

$$F1\ measure = \frac{2 \times precision \times recall}{precision+recall} \quad (8)$$

4. RESULTS AND ANALYSIS

4.1 Insights by subject matter expert

When looking at every question presented to students over the course of a semester, the subject matter expert identified the number of questions corresponding to Remember, Apply and Transfer as shown in Table 4. Just by labeling the course questions, the subject matter expert realized how misaligned the course's learning outcomes were with its assessment practices. A large emphasis on Apply questions was expected, but the dearth of Transfer questions was surprising. Of those 23 Transfer items, most were presented during the final exam.

Table 4 - Frequency of questions aligned to learning outcomes

Learning outcome	Frequency (number of questions)
Remember	62
Apply	131
Transfer	23

One of the stated learning outcomes of the course was to prepare students to flexibly transfer course content to novel problems and new situations. However, waiting until the final exam to present students with such opportunities denied them actionable feedback during the semester. In response to the pre-processing labeling efforts, the subject matter expert then added 42 new transfer questions throughout the course for the next semester.

4.2 Model reliability with subject matter expert

The objective of this implementation is to evaluate whether the trained model is able to predict the type of question (Remember, Apply or Transfer). Based on the trained model using questions from the undergraduate course, the testing questions from textbooks and online sources were passed through our model to determine the level of reliability of labeling new questions that were not generated by the subject matter expert. In our intended use case, the testing dataset would not need to be manually labeled. However, to determine the level of reliability of our labeling algorithms, the subject matter expert's manual labels served as a ground truth for the F1 measure calculations.

4.2.1 Parameter selection

We first determined the best set of parameters based on 10-fold cross validation of the training dataset. As there were 120 questions, 90% of the questions (108 questions) were used for training and 10% of the questions (12 questions) were used as a validation set. This process was done 10 times using 10 different bundles of the 120 questions. The best set of parameters were chosen based on a grid search for both ELM and SVM.

The parameters that were varied for ELM were:

1. Number of hidden nodes
2. Activation function (sigmoid / radial basis / hard-limit)

The parameters yielding the best results corresponded to 72 hidden nodes using hard-limit activation function.

The parameters that were varied for SVM were:

1. Kernel (sigmoid / radial basis)
2. Cost value
3. Gamma value

The parameters yielding the best results corresponded to sigmoid kernel, cost value = 1, gamma value = 0.26

4.2.2 Comparing four combinations

With respect to the F1 measure, calculations were done separately for the three labels. The mean of those calculations was then used as the algorithm's overall performance measure. With respect to ELM, the calculation was repeated 10 times because the initialization weights are randomly assigned in each iteration. The mean value of the F1 measure was taken.

Table 5 below shows the F1 measure values (for each individual class and overall F1 mean) for the four combinations. "R" refers to Remember, "A" refers to Apply, "T" refers to Transfer and "s.d." refers to standard deviation.

Table 5 - F1 measure values for four combinations

Combination	R	A	T	Mean	s.d.
1. TF-IDF with SVM	0.870	0.737	0.667	0.758	0.084
2. LDA with SVM	0.400	0.593	0.556	0.516	0.084
3. TF-IDF with ELM	0.926	0.815	0.840	0.860	0.048
4. LDA with ELM	0.467	0.520	0.647	0.545	0.076

TF-IDF with ELM achieved the highest mean F1 measure value and the lowest standard deviation – indicating that it was the most reliable combination. It can be seen that the Remember label yields the highest F1 values out of the three labels in Combination 3. In general, Remember-labeled questions are short, resulting in about four to five zero values in the TF-IDF vector of 10 columns that is passed as an input into the ELM. Hence, the algorithm identifies Remember-labeled questions very accurately due to their size.

The result of high reliability in using ELM is as expected because it has already been demonstrated that ELM outperforms SVM when comparing in terms of standard deviation of training and testing root-mean-square values, time taken, network complexity, as well as performance comparison in real medical diagnosis application [23]. On the other hand, although LDA has been shown to achieve higher performance as it groups words together in terms of topics instead of looking at combinations of individual words which may not link together, in the context of our work, TF-IDF outperforms LDA instead. This is because for LDA, the goal is to correctly assign each document (or question) to a class label in a reduced dimensional space [31]. However, in our corpus of questions, there are several technical terms involved, without any prior labeling of topics. Hence, LDA is not appropriate for our analysis.

5. CONCLUSIONS

Based on the comparison of our four algorithms, our most reliable model (TF-IDF with ELM) is able to accurately label new course questions for the undergraduate electrical and electronic engineering course with 0.86 reliability in terms of F1 measure. Any novice instructor who takes over this course in the future or teaching assistants tasked with refreshing the course assignments would be able to extract new questions from any external source and pass them to the algorithm to automatically label the questions as the original course coordinator would. This allows members of the course design team without a strong background in learning to make curriculum decisions regarding the alignment of the course's learning outcomes.

As discussed earlier, outcome-based learning environments facilitate transforming the model of instruction from instructor-centric and lecture-based to being more learner focused filled with a variety of activities and learning pathways. However, in learner-centered environments, assessment is still the key driver, and often the key inhibitor of learning [3]. If the assessments require shallow understanding, then learners calibrate their efforts to achieve this low bar. When assessments require deep understanding or great proficiency, learners are likely to put in more effortful practice.

In line with this assessment philosophy, our TF-IDF with ELM model is theoretically capable of matching any learning activity to any set of learning outcomes as long as the course designers or subject matter experts provide enough examples that are explicitly aligned to the intended learning outcomes when training the model. For the convenience of the subject matter expert in our context, we used a reduced version of Bloom's Taxonomy in this study. However, the final algorithm is capable of using the full Bloom's model, a different model, or a custom set of learning outcomes as its labeling framework.

Hence, with the high reliability of the prediction algorithm presented in our work, our process for calibrating the algorithm can be used in any academic or industrial setting to provide the right set of formative assessment opportunities to students (enhancing subject knowledge) or employees (professional development). Once the learning outcomes of activities are labeled reliably, it is then easier to think about how to engage learners in deliberate practice to reach those outcomes and develop their expertise. Once opportunities for deliberate practice that align to the course learning outcomes are implemented into a course, it becomes easier to think about how to align the feedback regarding those opportunities to support the development of domain expertise.

This work provides a first step at being able to regularly introduce learning activities that promote the development of adaptive expertise into a course by matching external sources of activities with the course's learning outcomes. Deliberate practice requires repetition that varies in ways that highlight the structural elements of a domain. Having a way to incorporate new sources of questions and problems into a course that align with the course's goals provides learners more opportunities for internalizing when to apply their domain specific skills and knowledge. Finally, our algorithm is potentially useful for designing courses to reach non-content-based learning outcomes, making policies that support constructive alignment, and evaluating course assessment of learning plans.

6. FUTURE WORK

Building off of our machine learning labeling work, we would like to explore constructing a new version of LDA that can be tailor-made to label questions. There are situations in which weightages given to words are the same, with different words representing those weightages. Similarly, the same words can have different weightages. We are keen to continue working on features based on word arrangement, word context and word order that affect weightage assignments. In addition, ELM can be enhanced by using kernels.

From the learning aspect, we would like to extend our question label categories to all six outcomes described in Bloom's Taxonomy and expand the model to label outcomes based on the types of sentences used in forum conversations and other collaborative learning activities. Eventually, we aim to determine the proficiency level of learners so we can put learning supports in place to guide their learning journeys. Ultimately, we wish to provide learners with learning activities and opportunities for deliberate practice embedded with actionable feedback to develop their adaptive expertise.

7. ACKNOWLEDGMENTS

This work was conducted within the Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme.

8. REFERENCES

[1] Krathwohl, D.R. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice*. 41, 4 (2002), 212-218. DOI= http://dx.doi.org/10.1207/s15430421tip4104_2

[2] Biggs, J. 1996. Enhancing teaching through constructive alignment. *Higher Education*. 32, 3 (1996), 347-364. DOI= http://dx.doi.org/10.1007/BF00138871

[3] Boud, D. 2010. Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*. 22, 2 (2010), 151-167. DOI= http://dx.doi.org/10.1080/713695728

[4] Boud, D. and Falchikov, N. 2006. Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*. 31, 4 (2006), 399-413. DOI= http://dx.doi.org/10.1080/02602930600679050

[5] Miller, G. E. 1990. The Assessment of Clinical Skills/Competence/Performance. *Academic Medicine*. 65, 9 (1990), S63-S67. DOI= http://dx.doi.org/10.1097/00001888-199009000-00045

[6] Wass, V. et al. 2001. Assessment of clinical competence. *The Lancet*. 357, 9260 (2001), 945-949. DOI= http://dx.doi.org/10.1016/S0140-6736(00)04221-5

[7] Hmelo-Silver, C.E. 2004. Problem-based learning: What and how do students learn? *Educational Psychology Review*. 16, 3 (2004). 235-266. DOI= http://dx.doi.org/10.1023/B:EDPR.0000034022.16470.f3

[8] Biggs, J. B. and Collis, K.F. 2014. *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcomes).* Academic Press.

[9] Crowe, A. et al. 2008. Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE-Life Sciences Education.* 7, 4 (2008), 368-381. DOI= http://dx.doi.org/10.1187/cbe.08-05-0024

[10] Ericsson, K.A. et al. 1993. The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review.* 100, 3 (1993), 363-406. DOI= http://dx.doi.org/10.1037/0033-295X.100.3.363

[11] Duckworth, A. L. et al. 2007. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*. 92, 6 (2007), 1087. DOI= http://dx.doi.org/10.1037/0022-3514.92.6.1087

[12] Schwartz D. L. et al. 2005. Efficiency and innovation in transfer. *Transfer of learning from a Modern Multidisciplinary Perspective*. Information Age Publishing. 1-51.

[13] Chi, M. T. 2006. Two approaches to the study of experts' characteristics. *The Cambridge Handbook of expertise and expert* performance. Cambridge University Press. 21-30.

[14] Black, P. and William, D. 1998. Assessment and Classroom Learning. *Assessment in Education Principles Policy and Practice.* 5, 1 (1998), 7-74. DOI= http://dx.doi.org/10.1080/0969595980050102

[15] Ritter, S. et al. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review.* 14, 2 (2007), 249-255. DOI= http://dx.doi.org/10.3758/BF03194060

[16] Fei, T. et al. 2003. Question Classification for E-learning by Artificial Neural Network. In *Proceedings of the 2003 Joint Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia* (Singapore, 2003), 1-5. DOI= http://dx.doi.org/10.1109/ICICS.2003.1292768

[17] Cheng, S. C. et al. 2005. Automatic Leveling System for E-Learning Examination Pool Using Entropy-Based Decision Tree. In *Advances in Web-Based Learning – ICWL 2005* (Hong Kong, 2005), 273-278. DOI= http://dx.doi.org/10.1007/11528043_27

[18] Jayakodi, K. et al. 2015. An Automatic Classifier for Exam Questions in Engineering: A Process for Bloom's Taxonomy. In *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (Zhuhai, China, 2015). DOI= https://dx.doi.org/10.1109/TALE.2015.7386043

[19] Haris, S. S. and Omar, N. 2015. Bloom's taxonomy question categorization using rules and N-gram approach. *Journal of Theoretical and Applied Information Technology.* 76, 3 (2015), 401-407.

[20] Yahya, A. A. et al. 2013. Analyzing the cognitive level of classroom questions using machine learning techniques. In *The 9th International Conference on Cognitive Science* (Kuching, Sarawak, Malaysia, 2013). 587-595. DOI= http://dx.doi.org/10.1016/j.sbspro.2013.10.277

[21] Abduljabbar, D. A. and 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 (2015), 447-455.

[22] Sangodiah, A. et al. 2014. A Review in Feature Extraction Approach in Question Classification Using Support Vector Machine. In *2014 IEEE International Conference on Control System, Computing and Engineering* (Penang, Malaysia, 2014), 536-541. DOI= http://dx.doi.org/10.1109/ICCSCE.2014.7072776

[23] Huang, G. B. et al. 2006. Extreme learning machine: Theory and applications. *Neurocomputing.* 70, 1-3 (2006), 489-501. DOI= http://dx.doi.org/10.1016/j.neucom.2005.12.126

[24] Trinity University Course Assessment and Outcomes: 2016 *https://inside.trinity.edu/collaborative/collaborative-grants/course-redesign-stipends/course-assessment-and-outcomes*. Accessed: 2017-02-24.

[25] Bernardi, R. *Term Frequency and Inverted Document Frequency.* University of Trento, Trentino.

[26] Blei, D. M. et al. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3 (2003), 993-1022.

[27] Huang, G. B. 2015. What are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle. *Cognitive Computation.* 7, 3 (2015), 263-278. DOI= http://dx.doi.org/10.1007/s12559-015-9333-0

[28] Weston, J. *Support Vector Machine (and Statistical Learning Theory).* NEC Labs America, Princeton.

[29] Chang, C. C. and Lin, C. J. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST).* 2, 3 (2011), 1-39. DOI= http://dx.doi.org/10.1145/1961189.1961199

[30] Santra, A. K. and Christy, C. J. 2012. Genetic Algorithm and Confusion Matrix for Document Clustering. *IJCSI International Journal of Computer Science Issues.* 9, 1 (2012), 322-328.

[31] Hu, D. J. 2009. *Latent Dirichlet Allocation for Text, Images, and Music*.