Journal of Applied Testing Technology, Vol 23(Special Issue 1), 53-71, 2022

Using Machine Learning to Predict Bloom's Taxonomy Level for Certification Exam Items

Alan D. Mead[1]* and Chenxuan Zhou[2]

[1]Chief Psychometrician, Certiverse, 4803 N. Milwaukee Avenue, Suite B, Unit 103, Chicago, IL 60630, USA; alan@certiverse.com
[2]Psychometrician, Certiverse, 4803 N. Milwaukee Avenue, Suite B, Unit 103, Chicago, IL 60630, USA; chenxuan@certiverse.com

Abstract

This study fit a Naïve Bayesian classifier to the words of exam items to predict the Bloom's taxonomy level of the items. We addressed five research questions, showing that reasonably good prediction of Bloom's level was possible, but accuracy varies across levels. In our study, performance for Level 2 was poor (Level 2 items were misclassified and other items were classified as Level 2), but the performance of the model in distinguishing Level 1 from all other levels was quite good. Applying a model developed on an IT certification exam domain to a more diverse set of items showed poor performance, suggesting that models may generalize poorly. Finally, we showed what features of items the classifier was using. Examples and implications for practice are discussed.

Keywords: Naïve Bayesian classifier, Bloom's Taxonomy, Machine Learning

1. Introduction

Bloom's Taxonomy (Bloom *et al.*, 1956; Anderson & Krathwohl, 2001) is a framework for categorizing the cognitive complexity of education and assessment materials, such as learning objectives and exam items. This taxonomy enjoys wide usage in the field of exam development. This paper briefly describes Bloom's taxonomy and how it may be applied to the exam development process, including pitfalls that should be avoided. We then sketch a rationale for automated classification into Bloom's taxonomy and turn to the main issue of predicting the Bloom's classification of certification exam items using a combination of Machine Learning (ML) models using features engineered from Natural Language Processing (NLP) models. The discussion addresses how this research is connected to the literature and explores additional practical implications for using ML and NLP can be used to make exam development more efficient.

2. Background and Literature Review

2.1 Bloom's Taxonomy

Bloom's Taxonomy was first created during the 1950s by Benjamin Bloom and his colleagues as a set of hierarchical models used to classify educational objectives in the cognitive, affective, and psychomotor domains (Bloom *et al.*, 1956). The cognitive domain is the most widely used domain for educators to assess the cognitive complexity of the education and assessment materials, such as learning objectives and exam questions. The taxonomy for the cognitive domain was more recently revised by Anderson

*Author for correspondence

Table 1. Revised Bloom's taxonomy cognitive domain levels

#	Level	Definition	Exemplar Verbs
6	Create	Put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure	adapt; build; change; choose; combine; compile; compose; construct; create; design; develop; discuss; elaborate; estimate; formulate; imagine; improve; invent; modify; plan; predict; propose; solve; suppose; test
5	Evaluate	Make judgments based on criteria and standards	appraise; assess; award; choose; compare; conclude; criticize; decide; deduct; defend; determine; disprove; estimate; evaluate; explain; influence; interpret; judge; justify; mark; measure; perceive; prioritize; prove; rate; recommend; rule on; select; support; value
4	Analyze	Break material into its constituent parts and determine how the parts relate to one another and to an overall structure or purpose	analyze; assume; categorize; classify; compare; conclude; contrast; discover; dissect; distinguish; divide; examine; function; infer; inspect; simplify; survey; theme
3	Apply	Carry out or use a procedure in a given situation	apply; build; choose; construct; develop; experiment with; identify; interview; make use of; model; organize; plan; select; solve; utilize
2	Understand	Construct meaning from instructional messages, including oral, written, and graphic communication	classify; compare; contrast; demonstrate; explain; extend; illustrate; infer; interpret; outline; relate; rephrase; show; summarize; translate
1	Remember	Retrieve relevant knowledge from long-term memory	choose; define; find; label; list; match; name; omit; recall; relate; select; show; spell; tell

Note. Based on Anderson *et al.* (2001).

and Krathwohl (2001), where the levels were renamed and slightly adjusted in the order of complexity. The two versions of the taxonomy are referred to as Bloom's taxonomy (Bloom *et al.*, 1956) and the revised Bloom's taxonomy (Anderson *et al.*, 2001).

Table 1 shows the six revised Bloom's taxonomy cognitive domain levels (Anderson *et al.*, 2001) from Remember at the lowest level, requiring simply retrieving the information from long-term memory, to Create at the highest, representing the ability to organize different elements in a new way or to propose alternative solutions. The revised Bloom's taxonomy levels were used in the current study to label the items, but our literature review will cover previous research using both the original and the revised Bloom's taxonomy levels.

The Bloom's taxonomy cognitive domains have been widely used as aid for educators to set learning objectives and to design the curriculum. The verbs in Table 1 are an indication of the kinds of verbs that would be used in learning objectives, although the classification into level is more holistic than merely based on the verb used. For example, the verb 'choose' is shown for levels 1, 3, 5 and 6, indicating that differing contexts could place a learning objective with the verb 'choose' at almost any level of the taxonomy.

From its inception Bloom's taxonomy has been meant to facilitate interchange regarding assessment: "…a common framework for classifying intended student learning outcomes could promote the exchange of test items, testing procedures, and ideas about testing. (p. xxvii, Anderson *et al.*, 2001)." The broad definition representing cognitive processes in general, instead of some classification limited to a specific content domain, allowed the taxonomy to be applied to various domains. On the other hand, being broadly defined also requires researchers and practitioners to relate the framework to specific tasks before applying to different disciplines. For example, previous work has been conducted to contextualize the taxonomy to biology (Crowe *et al.,* 2008), computer science (Thompson *et al.*, 2008), management (Athanassiou *et al.*, 2003), and medical courses (Nkanginieme, 1997). Contextualizing the taxonomy for a domain is a considerable amount of work. Although the methodology presented in this paper is radically different, our approach could be thought of as an automated process for contextualizing the taxonomy to any item pool.

Bloom's taxonomy has been used widely in exam development for multiple purposes. We identified at least three uses specific to exam development. First, the taxonomy can readily be applied to exam objectives and then again to the items written to assess those objectives to ensure that levels of cognitive complexity required in the exam questions match those set for the corresponding learning objectives.

In addition, good assessment practices should support learning, which foster students' understanding of the learning materials and facilitate their ability to integrate and apply what they have learned (Buckwalter *et al.,* 1981; Cilliers *et al.,* 2012; Jensen *et al.*, 2014). Thus, written items are usually preferred to engage higher-level cognitive processing, rather than simply requiring recall of information. For example, the National Board of Medical Examiners Item Writing Guide specifies "Each item should assess application of knowledge, not recall of an isolated fact (Billings *et al.*, 2020, pp. 29)." Thus, a second use of Bloom's Taxonomy is to avoid lower-level items and ensure adequate numbers of items assessing higher-order cognitive functions.

Finally, Bloom's levels are often controlled to ensure challenging exam content based on an assumption that higher Bloom's items are more difficult (or that items of lower levels are trivially easy). The evidence for an association between Bloom's level and item difficulty is mixed. Tan and Othman (2013) found that Bloom's taxonomy was only weakly related to item difficulty. The largest predictor in a regression analysis predicting Rasch item difficulty of physics items was the type of item (multiple-choice vs. constructed response; Mesic & Muratovic, 2011). On the other hand, two studies of TIMSS item difficulty found Bloom's level to be substantially related to item difficulty. Rosca (2004) reported a partial correlation of 0.499 (after partialling out the other predictors). Sinharay (2016) found the first three levels of Bloom's taxonomy to be the second most important splitting variable in building a classification and regression tree. Thus, although Bloom's level may be related to item difficulty, the relationship varies and is unlikely to be uniformly strong. Nonetheless, if Bloom's level contributes to predictions of item difficulty, then an automated methodology for classifying items into the taxonomy could improve the accuracy and efficiency of exam development by improving the targeting of the desired level of item difficulty.

There are at least two obstacles to effectively using Bloom's taxonomy in exam development: the amount of specialized labor required, and the reliability of classification (Karpen & Welch, 2016). The verbs shown in Table 1 may be helpful in classifying learning objectives, but those verbs have, at best, a weak association with the cognitive level required to solve an item. Also, as previously discussed, the verbs are not unique to a level; instead, a holistic judgement about the item is needed. As a result, classifying items into Bloom's taxonomy requires subject matter expertise. A Subject Matter Expert (SME) has the knowledge needed to understand the process of solving an item, but they are unlikely to understand or appreciate Bloom's taxonomy. Therefore, test developers face the choice of training item authors to classify items into the appropriate levels or employing a SME reviewer to perform the Bloom's classification.

Another problem with using Bloom's taxonomy is that the classification of exam items can have low reliability. Karpen and Welch (2016) surveyed 21 pharmacy faculty asking them to classify 126 items chosen to be evenly distributed across the six levels of Bloom's taxonomy. Overall, the accuracy was 46.0% but they provided a confusion matrix (described in the next section of the paper) showing that while Bloom's level 1 items were

correctly classified 95.2% of the time, classifications into other levels were far less accurate, between 28.6% and 47.6%. The present authors have independently coded pools of items for this line of research, and our agreement is typically in the 65% to 75% range. Furthermore, this is unsurprising because our experience during classification has been that some items are difficult to classify. If we felt uncertainty in our own classifications, it would be unsurprising if different evaluators reached different conclusions about items.

Two studies have reported better agreement for collapsed taxonomies (Karpen & Welch, 2016; Plack *et al.*, 2007). The Plack *et al.,* study collapsed level 1+2, 3+4, and 5+6 and reported an average accuracy of 88.9% (ranging from 78.2% to 100.0%) using these levels. The Karpen and Welch study examined a taxonomy with three collapsed levels: 1, 2+3, 4+5+6 and reported an accuracy of 81.7% using these levels.

In our experience, using a collapsed taxonomy is not uncommon, however the traditional solutions to improving the reliability of a classification task are to train judges and to have multiple judges. Both strategies exacerbate the labor required to perform the classification of large item pools into Bloom's taxonomy. Based on past machine learning studies, a statistical methodology is likely to be both extremely replicable and achieve an accuracy comparable to the results for collapsed categories.

As previously stated, this paper presents a machine learning method for predicting Bloom's level from the words of exam items. Although there are many machine learning models, psychometricians will appreciate that the method presented in the paper can be applied to any item pool in a way similar to IRT models and specifically without reference to large domain-specific corpus required by many machine learning models applied to natural language. The next section describes our approach and summarizes prior machine learning research on automated prediction of Bloom's taxonomy and presents our research questions.

2.2 Machine Learning using Naïve Bayesian Classification

A multinomial Naïve Bayes Classifier (NBC) model was selected for this project for three reasons. First, because the model is simple and easily applicable. It will be easy for the current authors, and the reader, to apply this methodology to different item pools in approximately the same was that IRT is easy to apply to exams of very different domains. Second, considerable research has shown that the NBC has considerable practical utility (Orrù *et al.,* 2020; Stephens *et al.,* 2018). The authors have routinely observed high degrees of predictive accuracy. And finally, some machine learning models are notoriously difficult to interpret. One of the goals we had for this research was to understand the automated classifications. In this regard, the parametric form of the NBC model (see below) is fairly easy to examine and provided considerable insight into how our models predicted Bloom's levels (see Tables 5, 6 and 7 in the results section).

The foundation for NBC model is Bayes' theorem which describes the probability of an event in relation to prior knowledge of the event (the prior) and the observed occurrence of conditions related to the event (the evidence). In classification work, we are interested in $P(c|d)$, the probability of a document d (e.g., an exam item) belongs to class c (e.g., a Bloom's level). Classification using the NBC model uses a "bag of words" approach (Jurafsky & Martin, 2000; Manning *et al.*, 2009) where each document is broken into a set of unique tokens, $t_1,t_2,...,t_k$ (syntax is ignored), and using Bayes' theorem the probability of interest becomes, $P(c|t_1,t_2,...,t_k)$, the probability of class c given the co-occurrence of these tokens:

$$P(c|d)=P(c|t_1,t_2,...,t_k)=\frac{P(t_1,t_2,...,t_k|c)P(c)}{P(t_1,t_2,...,t_k)} \quad (1)$$

where $P(t_1,t_2,...,t_k|c)$ is the conditional probability of tokens, $t_1,t_2,...,t_k$, occurring in a document we know to be from class c. $P(c)$ is the prior probability of a document belonging to class c. Finally, $P(t_1,t_2,...,t_k)$ is the probability of tokens in the data.

Applying the chain rule for repeated application of the conditional probability, Equation (1) becomes:

$$\breve{}\breve{}\breve{} \quad \frac{P(t_1|t_2,...,t_k,c)P(t_2|t_3,...,t_k,c)...P(t_{k-1}|t_k,c)P(t_k\breve{}\breve{}\breve{}}{P(t_1,t_2,...,t_k)} \quad (2)$$

Note that when predicting, the denominator is identical for all classes and thus practical implementations ignore

this term and calculate based on values proportional to the probability in Equation (2).

The term "naïve" signals an assumption that all the terms in a document *d* are independent, given the category *c*. Under this assumption:

$$P(t_1|t_2,...,t_k,c)=P(t_1|c) \quad (3)$$

This assumption is plainly incorrect, but makes the calculation of the NBC feasible, because most practical applications would not provide the data needed to calculate these terms (i.e., words do not co-occur in documents) and estimates from even sizeable samples would be too small to be accurate (extremely large samples would be needed to estimate all interactions well). Research has demonstrated that the NBC has considerable predictive power despite this assumption (Stephens *et al.*, 2018; Orrù *et al.*, 2020).

Applying Equation (3) in each of the conditional probability terms in Equation (2), the probability of a document *d* being in class c, $P(c|d)$ is computed as:

$$P(c|d) \propto P(c) \prod_{1\le k\le nd} P(t_k|c) \quad (4)$$

given that the document belonging to class *c*, and *nd* is the number of tokens in a document used for classification.

The Naïve Bayes Classifier combines the Naïve Bayes probability model with a decision rule. The maximum a posteriori, or MAP, decision rule is a commonly used one where the document is assigned to the class that is most probable. According to Manning *et al.* (2009), the most probable class, c_{map}, can be estimated as:

$$c_{map} = \arg\max_{c\in C} \hat{P}(c|d) = \arg\max_{c\in C} \hat{P}(c) \prod_{1\le k\le nd} \hat{P}(t_k|c) \quad (5)$$

In Equation (5), many conditional probabilities are multiplied, which can result in arithmetic underflow. Therefore, the computation in most NBC implementations is performed by adding logarithms of the probabilities (Manning *et al.*, 2009). Thus, applying $log(xy)=log(x)+log(y)$ to Equation (5), c_{map} is estimated as:

$$c_{map} = \arg\max_{c\in C} [\log \hat{P}(c) + \sum_{1\le k\le nd} \log \hat{P}(t_k|c)] \quad (6)$$

In text classification, the estimated terms $\hat{P}(c)$ and $\hat{P}(t_k|c)$ are computed as:

$$\hat{P}(c) = \frac{N_c}{N} \quad (7)$$

where, N_c is the number of documents in class c and *N* is the total number of documents, and:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t'\in V} T_{ct'}} \quad (8)$$

where, T_{ct} is the number of occurrences of term *t* in training documents belonging to class *c* and *t* has to be part of the vocabulary, *V*, in the training documents.

In summary, the NBC model is a simple model, easily calculated, and based on two intuitive assumptions: that high base-rate classes should be predicted more often than low base-rate classes, and that having previously seen a token associated with a Bloom's level should increase the likelihood of predicting that class for a new item.

Evaluation

The performance of a predictive classifier is commonly evaluated using the four indices: Accuracy, Precision, Recall, and F1. In order to define these indices, the terms True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) must be introduced: TP refers to the number of cases predicted to be a given class that are actually in that class, whereas FP is the number of cases incorrectly predicted to be a class. TN is the number of cases correctly predicted to not being in a class. Lastly, FN is the number of cases belonging to a given class that the model classified incorrectly (as some other class).

The model Accuracy measures the number of correctly classified documents in proportion to the total number of items, computed as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

Precision measures the ability of a classification model to identify the correct classification while avoiding false positives:

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

Recall measures the ability of a classification model to identify the correct classification while avoiding false negatives:

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

Within a single study when trade-offs are made between false positives and false negatives, then as precision increases, recall would decrease. There are instances where we favor maximizing either precision or recall, but for instances where an optimal balance of both is of interest the F1 measure combines the two indices as the harmonic mean of precision and recall, which differs from a simple average of the two in that it punishes extreme values in either index:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

Confusion Matrix

A final method of evaluation, analogous to a scatterplot in correlation analysis or a residual plot in regression analysis, is the "confusion matrix," which is a matrix showing the accuracy as tallies of correct and incorrect classifications. The name confusion matrix emphasizes the value of this display to understand where the model is making classification mistakes (e.g., classifying a Bloom's level 1 item as some other level, such as level 3). Perfect prediction is shown by a diagonal confusion matrix with zeros in the off diagonal. For typical (imperfect) results, the location of non-zero values indicates where a class was predicted to be another class and the frequency of these misclassifications. Examples of confusion matrices can be seen in the results section, along with our interpretations.

The next section describes how the NBC model is applied to textual data using natural language processing (NLP) and then reviews prior studies using ML and NLP.

2.3 Natural Language Processing, Feature Engineering, and Prior Machine Learning Studies

Although the NBC model presented in the previous section may seem complete, in order to maximize prediction in practice, the words of items need to be pre-processed in a procedure called "feature engineering" where the predictor values are constructed from noisy input data using domain-specific knowledge to reduce noise and boost signal in the predictor features. This procedure is analogous to scoring items in conventional psychometric analyses. In the processing of natural language, this step of feature engineering is more complex and even more important. For example, the word forms of the lemma "analyze" should all probably be counted together as one token, which is commonly achieved by "stemming" the words (Porter, 1980; e.g., "Analyze," "analyzes," "analyzed," and "Analyzing" are all stemmed to "analyz"; see Jurafsky & Martin, 2000 for more information). We say "probably" because there may be unusual circumstances in which stemming is unhelpful, and stemming works primarily on regular wordforms and verbs (e.g., the word "was" is stemmed to "wa", "geese" is stemmed to "gees"). In an additional step, common, low-information words, called "stopwords," are typically removed (i.e., ignored). Finally, spelling, punctuation, capitalization, and any text or font styles are typically removed. Thus, an item like: "Which ONE of the following rules apply to case 12?" would be analyzed with the following tokens: "follow", "rule", "appli", "case." The tallies that define the estimated condition probabilities would be for those tokens, and it is essential that identical feature engineering occurs during model training and later prediction. Again, an analogy to conventional psychometrics would be the importance of scoring items with the same key and procedures before a psychometric analysis and later when using the scores. Just as scoring items on an exam requires an answer key, effective feature engineering requires an understanding of the domain. In this study, we applied generic, English-language methods that would not require tuning to the domain.

Prior research has attempted to automate the process of classifying exam materials based on Bloom's taxonomy levels. Chang and Chung (2009) presented an online test system that extracted all the verbs from an input question,

matched them to an existing pool of keywords, each associated with one or more Bloom's levels, and assigned a class to the question based on the matching results. Their results showed a 75% overall accuracy in classifying items of Bloom's level 1. Similarly, Omar *et al.* (2012) proposed a keyword matching and rule-based approach in tackling item classification. Haris and Omar (2015) further refined this rule-based approach by analyzing the speech structure of the exam items and developing a set of rules to distinguish cognitive levels based on a combination of part-of speech, regular expressions, and predefined keywords. They also implemented a hybrid approach where the rule-based classification was integrated with statistical learning approach, which produced 86% F1 averaged across levels.

Researchers have also implemented machine learning techniques in classifying exam materials, and various machine learning models have been utilized, including K-Nearest Neighbors (KNN), artificial neural networks (ANN), Support Vector Machines (SVM), and Naïve Bayes (NB). Among works dedicated to classifying exam questions into Bloom's cognitive levels, Yahya and Osaman (2011) used Support Vector Machines to classify 272 open-ended questions, and the resulting model had satisfactory accuracy (ranged from 83% to 90%) and precision (ranged from 50% to 100%) in the cross-validation sample, but low recall (ranged from 9% to 64%). Yusof and Hui (2010) proposed an ANN model with various feature reduction methods, and the resulting precision ranged from 61% to 66% in the cross-validation question sample. Yahya *et al.* (2013) compared the effectiveness of KNN, SVM, and NB models in scenarios where different number of terms were used in question representation. They concluded that, in terms of classification accuracy and typical F1, the performance of the KNN and NB machine learning models were comparable, and that of the SVM was slightly better than the other two. However, the performance of SVM varied the most when different number of terms were used for question representation.

The first research question simply addressed the degree of accuracy for an ML model predicting Bloom's level of exam items. There are two primary reasons why additional research is required, beyond the studies summarized above. First, it is unclear the degree to which these prior studies used unedited items. When we apply this method to our exams, we do not plan on any manual editing of the items (as was done by Mohammad & Omar, 2020; see the Method section) before automated prediction of Bloom's Level. The second reason for additional research was to establish the degree of accuracy in a new exam domain, similar to how local criterion-related validation is recommended for employee selection tests. Besides differences in wording usage across domains, there may also be differences in the base rates of the levels, which affects predictive accuracy.

Research Question 1:
How well does the ML model (Naïve Bayesian classifier) predict Bloom's taxonomy levels of the test items?

The second research question addressed the degree of variability across levels. It stands to reason that prediction will be easier for lower and upper Bloom's Levels; Level 1 and 6 items can only be misclassified in a single direction, while items of intermediate levels can be misclassified in two ways. And variation in base rates of Bloom's levels in an item pool would also contribute to potentially large variation in the predictive accuracy of any statistical model.

Research Question 2:
Does the ML model have similar predictive accuracy in all Bloom's taxonomy levels?

As described earlier, sometimes the item writing practice prefers higher level items over level 1 items, while making no distinctions among higher level items (see Billings *et al.*, 2020). Given previous finding on improved classification reliability in collapsing taxonomies (see Karpen & Welch, 2016; Plack *et al.*, 2007), it is suggested that simply classifying items into Remember vs. higher levels may have advantage in these scenarios. However, machine learning models are rarely investigated with various level-collapsing conditions. Therefore, we designed Research Question 3 to examine the performance of a binary NBC model, and to further compare it to that of the multinomial NBC model.

Research Question 3:
How well does the ML model distinguish items of Bloom's taxonomy level 1 from items of level 2 and above?

Previous research has established the effectiveness of various machine learning models in classifying items using Bloom's taxonomy. Most of the studies have adopted the common approach to split a sample of items from one content-domain into training and cross-validation

subsets, but there is a lack of work demonstrating the performance of ML models when applied to items of a different domain. Although the context-specific nature of the taxonomy implies that a model effectively classifying items from one content domain may not transfer to items from another domain, this is rarely tested in empirical studies. Therefore, we address the following research question:

Research Question 4:
How well does the model fit in one domain cross-validate in another domain?

Finally, to better understand how ML models contextualize Bloom's levels for our domain and to better understand how the predictions were made, we were interested in interpreting our ML models. Some literature on Bloom's levels emphasizes the role of verbs in classifying items, but our experience suggests that this linkage is likely to be stronger for learning objectives or constructed response items than for multiple-choice exam items. Therefore, our last research question was:

Research Question 5:
What aspects of items influence the predictions of the model?

3. Method

3.1 Item Samples

The units of analysis in this study are items and two samples of items were used. Most of our analyses used a sample of 141 English-language, four-option multiple-choice questions extracted from online practice exams which assessed information technology related topics. Only items stated in the form of a question were used. This sample is referred to as the MCQ dataset. This sample is probably on the smaller end of the range of operational item pools.

The second sample was used to address RQ4 about the crossvalidity of the fitted NBC model and was provided in the supplementary materials of the Mohammad and Omar (2020) study, and consisted of 141 English-language, open-ended questions with those authors' classification of Bloom's taxonomical level for each item. (It is a coincidence that both samples of items had 141 items.) A careful review of this dataset revealed that the "items" were actually a mix of learning objectives and edited items. The items were edited to remove any "scenario" or other extraneous words from the item. For example, the literal text of a Level 2 item was: "Explain what is happening... " These items had heterogeneous response formats (multiple-choice and constructed response) and covered various educational content domains, including computer science, English, mathematics, history, etc. These items will be referred to as the cross-domain dataset.

3.2 Procedure

Rater Judgment
Items of the MCQ dataset were classified by a single rater into the six revised Bloom's taxonomy levels based on the words of the item. After classification, due to a lack of items in the upper levels, levels 4, 5 and 6 were collapsed into a single class. The NBC was trained to predict items in the four classes: Bloom's level 1 (Remember), level 2 (Understand), level 3 (Apply), and level 4 and above (Analyze, Evaluate, and Create).

To assess cross-domain classification, we used the items of the cross-domain dataset and the class labels provided by Mohammad and Omar (2020), which used the original 1956 version of Bloom's taxonomy. However, because items labeled as level 4 or above were collapsed into a single class, we determined that these class labels were fully consistent with the NBC model trained on the MCQ datasets (using the revised Bloom's taxonomy). This model was used to predict the Bloom's taxonomy levels of these cross-domain items. In addition, both authors of the current study also independently classified about 60% of items in the cross-domain dataset. The performance of our manual classification was compared to that of the NBC model.

Feature Extraction
Some stems of the MCQ dataset items were composed just of a single question, and others had one or two preceding background/scenario sentences. We parsed the stems into "scenario" and "question" parts and only the "question" was used in the later analysis (any additional words in the stem and the words of the response options were ignored). The "question part" of the stem was the full sentence or question phrase that ended in a question mark. For example, the question part of the stem is italicized in this example: "The ACME webserver runs on a Linux host in an off-premises data center. If some

webpages work and others are giving errors, *which of the following is the best first step in trouble-shooting the website errors?*" This was done (using Excel functions) to exclude scenario/background words that are likely to be used in ways that may not generalize well; this procedure was adopted because it improved our predictive results. A similar step had already been conducted for the items of the cross-domain dataset.

Using R packages 'tm' and 'SnowballC,' the words of each item were normalized by the following steps: converting to lower case; removing punctuation, numbers, and stopwords (from the default 'tm' English-language stopword list); stemming; and tokenization. The dataset was converted to a document-term matrix (DTM), where each document is in a row and the columns representing terms appearing across all items in the corpus. The individual cells of the matrix were 1 if the term appeared one or more times in the document and zero otherwise.

Classification

Using R package 'caret' the items were randomly split into 70% training and 30% testing samples. To ensure similar base rates in the training and testing samples, the splitting procedure stratified on Bloom's classifications prior to sampling. Using the naiveBayes() function of the 'e1071' package, the NBC model was fit to the training data using the default settings. Then the predict() function was used (with threshold = 0.001 and eps = 0) to predict the Bloom's level for both the training and testing samples. As in other prediction contexts, it is to be expected that the NBC model would be highly accurate in the training sample, whereas the performance in the testing sample provides a lower but more 'honest' estimate of the expected accuracy in future samples drawn from the same population.

4. Results

The MCQ dataset was split into a training sample with 101 questions and a cross-validation sample with 40 questions. The split was stratified on the labeled Bloom's level, but otherwise random. As shown in Table 2, the base rates for items in each Bloom's level were similar in both samples with about 45% of the questions from level 1 (Remember), about 17% level 2 (Understand), about 25% level 3 (Apply), and about 13% level 4 or higher (Analyze, Evaluate, and Create). The feature extraction process in the training sample resulted in a total of 357 unique terms.

RQ1. Research question 1 addresses overall accuracy: *How well does the ML model (Naïve Bayesian classifier) predict Bloom's taxonomy levels of the test items?* The overall performance of the NBC model is shown in Table 2. The training sample results are shown on the left and the cross-validation sample on the right. Prediction is almost perfect in the training sample, but these results overstate the cross-validated accuracy. In the cross-validation sample, the model had an overall accuracy = 0.775, F1 = 0.762, with balanced Precision = 0.756 and Recall = 0.775. These measures are good and substantially better than chance (0.25) or the "no information rate" (0.45). The Kappa level of 0.67 would be interpreted as "substantial" agreement for human judgments (Landis & Koch, 1977). However, performance varied across cognitive levels.

RQ2. Research question 2 addressed variation in prediction across levels: *Does the ML model have similar predictive accuracy in all Bloom's taxonomy levels?* As shown in Table 2, the model performance was the best for level 1 (F1 = 0.889) and level 4 or higher (F1 = 0.909) items, satisfactory for level 3 (F1 = 0.762), and not very good for level 2 (F1 = 0.333). As shown in the confusion matrix, the model incorrectly classified Level 2 items as belonging to other levels, as well as other levels as level 2. Only two of the seven Level 2 items were classified correctly (29%), and only two of the five items classified as Level 2 were actually Level 2 (40%).

RQ3. Research question 3 was: *How well does the ML model distinguish items of Bloom's taxonomy level 1 from items of level 2 and above?* To investigate the effectiveness of NBC in distinguishing low level (i.e., Remember) items from items greater than level 1, a two-level NBC was tested in the MCQ dataset all items labeled level 2 and higher were collapsed into a single class. The base rates for level 1 vs. level 2 and above items were around 0.45 and 0.55 in both the training and the cross-validation sample. As shown in Table 3, in cross-validation the two-level NBC model had an accuracy of 0.854 and F1-measure of 0.842. Clearly, this model has a high degree of accuracy in this domain distinguishing level 1 items from those of Bloom's level 2 and above.

RQ4. *How well does the model fit to items of one domain cross-validate in another domain?* To address RQ4, the four-level NBC trained in the MCQ dataset was applied to predict the levels of the 141 items in the cross-domain dataset. As shown in Table 4, the cross-domain sample has a different set of class base rates compared to the

Table 2. Results of the four-level NBC in the MCQ dataset

	Training sample					Cross-validation sample				
Bloom's level	**1**	**2**	**3**	**4**	**Overall**	**1**	**2**	**3**	**4**	**Overall**
Descriptives										
number of items	45	17	25	14	101	18	7	10	5	40
base rate	0.446	0.168	0.248	0.139	1.000	0.450	0.175	0.250	0.125	1.000
Confusion matrix *(rows = predicted, columns = actual)*										
Level 1	44	0	0	0	44	16	1	1	0	18
Level 2	0	17	0	0	17	2	2	1	0	5
Level 3	0	0	24	0	24	0	3	8	0	11
Level 4	1	0	1	14	16	0	1	0	5	6
Performance by class										
Precision	1.000	1.000	1.000	0.875	0.983	0.889	0.400	0.727	0.833	0.756
Recall	0.978	1.000	0.960	1.000	0.980	0.889	0.286	0.800	1.000	0.775
F1-measure	0.978	1.000	0.960	1.000	0.980	0.889	0.333	0.762	0.909	0.762
Overall model performance										
Accuracy	0.980					0.775				
95% CI of Accuracy	(0.930, 0.998)					(0.616, 0.892)				
NIR	0.446					0.450				
p (Accuracy > NIR)	0.000					0.000				
Kappa	0.972					0.673				

Note. Bloom's level: 1 = Remember, 2 = Understand; 3 = Apply; and 4 = collapsed levels Analyze, Evaluate, and Create. The Overall Precision, Recall, and FI-measure statistics were calculated as the weighted average of the class statistics; NIR = no information rate, which is taken to be the largest class percentage in the dataset; p (Accuracy > NIR) = the p-value of the one-sided test that the accuracy is significantly higher than NIR; Kappa = Cohen's Kappa coefficient which measures the agreement between predicted and observed classes.

MCQ dataset. Specifically, most of the items (about 54.6%) in this sample were of level 4 and higher, while the largest class proportion of the MCQ sample was about 45% for level 1 items. The NBC model showed a lack of effectiveness in predicting item levels in the cross-domain sample, with overall accuracy = 0.213 and F1 = 0.218.

Reliability of classification into Bloom's taxonomy. To further investigate the poor generalization to the cross-domain pool of items, we estimated the reliability of the item classifications in the cross-domain sample. Since reliability limits validity, if these classifications are unstable, they are unlikely to be cross-validated. The two

Table 3. Results of the two-level NBC in the MCQ dataset

	Training sample		Cross-validation sample	
Bloom's level	**Level 1**	**Level 2 and above**	**Level 1**	**Level 2 and above**
Descriptives				
number of items	44	56	19	22
base rate	0.440	0.560	0.463	0.537
Confusion matrix *(rows = predicted, columns = actual)*				
Level 1	43	0	16	3
Level 2 and above	1	56	3	19
Performance by class				
Precision	1		0.842	
Recall	0.977		0.842	
F1-measure	0.989		0.842	
Overall model performance				
Accuracy	0.99		0.854	
95% CI of Accuracy	(0.946, 1.000)		(0.708, 0.944)	
NIR	0.560		0.537	
p (Accuracy > NIR)	0.000		0.000	
Kappa	0.980		0.706	

Note. Bloom's levels were grouped into two classes: Level 1 represents the Knowledge, or Remember, level, while Level 2 and above includes levels Understand, Apply, Analyze, Evaluation, and Create. NIR = no information rate, which is taken to be the largest class percentage in the dataset; p (Accuracy > NIR) = the p-value of the one-sided test that the accuracy is significantly higher than NIR; Kappa = Cohen's Kappa coefficient which measures the agreement between predicted and observed classes

authors independently re-classified about 60% of the items in the cross-domain sample, and these classifications were compared to the original labels provided by Mohammed and Omar (2020). Our classifications agreed with those of provided by Mohammed and Omar about 61%. Among the 49 items both authors classified, 24 (49%) were an exact match. The Pearson correlation between the authors' classifications was 0.68, while those between either author and the original labels were 0.79. Interpreted as a reliability estimate, this degree of reliability is modest, but comparable to other studies of applying Bloom's taxonomy (and might be slightly conservative as compared to a similar analysis using Spearman correlations). However, this level of reliability can only explain a small degree of

Table 4. Results of applying the four-level NBC in the cross-domain dataset

Cross-domain validation sample					
Bloom's level	**1**	**2**	**3**	**4**	**Overall**
Descriptives					
number of items	26	23	15	77	141
base rate	0.184	0.163	0.106	0.546	1.000
Confusion matrix *(rows = predicted, columns = actual)*					
Level 1	6	4	1	25	36
Level 2	2	3	2	15	22
Level 3	10	9	7	23	49
Level 4	8	7	5	14	34
Performance by class					
Precision	0.167	0.136	0.143	0.412	0.293
Recall	0.231	0.130	0.467	0.182	0.213
F1-measure	0.194	0.133	0.219	0.252	0.218
Overall model performance					
Accuracy	0.213				
95% CI of Accuracy	(0.148, 0.290)				
NIR	0.546				
p (Accuracy > NIR)	1.000				
Kappa	-0.038				

the lack of generalization to the cross-domain dataset items.

RQ5. Our final research question addressed how predictions are made by the NBC model. Many past applications of Bloom's taxonomy have emphasized the role of verbs. Tables 5 and 6 show detailed numeric examples of predictions for two sample items. Table 7 shows the tokens that were most indicative of the four classes.

Table 5 shows how example item 63 would be classified. This item used the verb "requir" and also had the terms "creat", "new", and "plan". All other words in the item were stopwords. We had classified the item as level 1 because the term "plan" was processed from "Plan" which is a proper noun in the context, and the question asks about the basic settings, or requirements, of the notion "Plan." The NBC model fitting shows that all four of these words were present in level 1 items in the training sample, and generally *more* common *or absent* for higher levels. Although "requir" or "creat" were more common in all levels above 1, "requir" did not appear in any level 2 or 3 items, "new" did not appear in any level 3+ items, and

Table 5. An example of the prediction calculations; predicting level for item 63.

Class	c1	c2	c3	c4
Prior probability	$\hat{P}(c_1)$	$\hat{P}(c_2)$	$\hat{P}(c_3)$	$\hat{P}(c_4)$
	0.4455	0.1683	0.2475	0.1386
Likelihood of *t* given *c* / **Term**	$\hat{P}(t\|c_1)$	$\hat{P}(t\|c_2)$	$\hat{P}(t\|c_3)$	$\hat{P}(t\|c_4)$
requir	0.0444	0.0000	0.0000	0.1429
creat	0.0889	0.1176	0.1200	0.1429
new	0.0222	0.0588	0.0000	0.0000
plan	0.0222	0.0000	0.0000	0.0000
Posterior probability	$\hat{P}(c_1\|d)$	$\hat{P}(c_2\|d)$	$\hat{\smile}_3$	$\hat{P}(c_4\|d)$
	8.6664E-07	0	0	0

Note: Item 63 was classified as L1 by the authors and predicted as L1 by these calculations. Actual implementation details vary slightly to improve numerical precision (see the R code in e1071.predict()).

Table 6. A second example of the prediction calculations; predicting level for item 80.

Class	c1	c2	c3	c4
Prior probability	$\hat{P}(c_1)$	$\hat{P}(c_2)$	$\hat{P}(c_3)$	$\hat{P}(c_4)$
	0.4455	0.1683	0.2475	0.1386
Likelihood of *t* given *c* / **Term**	$\hat{P}(t\|c_1)$	$\hat{P}(t\|c_2)$	$\hat{P}(t\|c_3)$	$\hat{P}(t\|c_4)$
workflow	0.0222	0.1176	0.0800	0.1429
adjust	0.0222	0.0588	0.0000	0.0000
audio	0.0222	0.0000	0.0000	0.0000
transit	0.0222	0.0000	0.0000	0.0000
Posterior probability	$\hat{P}(c_1\|d)$	$\hat{P}(c_2\|d)$	$\hat{\smile}_3$	$\hat{P}(c_4\|d)$
	1.0864E-07	0	0	0

Note: Item 80 was classified by the authors as L2 and predicted to be L1 by these calculations.

Table 7. Terms with highest conditional probability for each class

c1		c2		c3		c4	
term	$\hat{P}(t\|c_1)$	**term**	$\hat{P}(t\|c_2)$	**term**	$\hat{P}(t\|c_3)$	**term**	$\hat{P}(c_4\|d)$
follow	0.7111	follow	0.2353	use	0.2000	creat	0.1429
option	0.4000	choos	0.2353	analyst	0.2000	workflow	0.1429
use	0.3778	campaign	0.1765	report	0.2000	segment	0.1429
user	0.2000	action	0.1176	campaign	0.1200	express	0.1429
box	0.2000	creat	0.1176	creat	0.1200	requir	0.1429
dialog	0.2000	data	0.1176	data	0.1200	result	0.1429
panel	0.2000	can	0.1176	user	0.1200	shown	0.1429
imag	0.1333	xxxx	0.1176	busi	0.1200		
layer	0.1333	purpos	0.1176	practition	0.1200		
order	0.1111	workflow	0.1176	best	0.1200		
clip	0.1111	edit	0.1176	task	0.1200		
take	0.1111	form	0.1176				
action	0.1111	method	0.1176				
work	0.1111	nondestruct	0.1176				
select	0.1111	set	0.1176				
tool	0.1111	three	0.1176				
		two	0.1176				

Note: These terms are stemmed, as described in the text.

"plan" did not appear in any items above level 1. Thus, the maximum probability was associated with level 1, which was chosen as the predicted value and the zero conditional probabilities in Table 5 drive the classification.

The calculations in Table 5 (and 6) show zero probabilities where the word did not appear in any item for a level, but the e1071 package actually compares the probability to a parameter, 'eps' (short for epsilon-range, 0 by default) and, when the probability is less than or equal to 'eps,' substitutes the parameter value 'threshold' (0.001, by default). Using these values, classes 2-4 have non-zero values, but class 1 remains the predicted class because the 'threshold' value of 0.001 is very small. For example, the conditional probability of 'requir' is somewhat larger (0.1429 for class 4 and 0.0444 for class 1; 3.2 times larger) but the threshold value of 0.001 for 'new' in class 4 is 22.2 times smaller than the conditional probability for class 1 (0.0222).

As a second example, show in in Table 6, item 80 used the verb "adjusts" and also has the terms "workflow," "audio," and "transition." It was labeled as level 2 because it assesses the workflow of performing a task, which requires understanding the processes, but was misclassified as level 1 by the NBC model. The calculations are clear:

although the term "workflow" was much more likely to be associated with levels above level 1 (3.6 to 6.4 times more likely) the other terms, particularly "audio" and "transition" did not appear in higher level items, leading to the maximum likelihood for classification as level 1.

Table 7 shows the tokens with the highest conditional probabilities (>0.10) for each class. We observe three interesting patterns. First, "follow" and "option" have very high conditional probabilities in class 1 because many level 1 questions use "[W]hich of the following options…". Second, "creat" shows up in the top lists for all classes except for c1. Finally, there is a trend to shorter lists with lower top conditional probabilities for higher levels, which probably indicates that the vocabulary of level 3-6 items (in classes c3 and c4) is more diverse. If so, this reinforces the observation that much of the predictive power of the NBC model comes from zero probabilities.

It can be hard to determine whether a token was used as a verb in the original question. For example, the word "follow" is usually a verb, but the token "follow" shown in Table 7 was rarely used as a verb. Our interpretation of Table 7 is that while verbs are present, they do not dominate, nor do they align particularly well with the exemplar verbs shown in Table 1. Thus, one summary of our findings for RQ5 is that the NBC model finds tokens characteristic of Bloom's levels to contextualize Bloom's taxonomy for a domain, and in a highly contextualized model, nouns and other parts of speech play a significant role; also, zero conditional probabilities seem to play a key role by excluding levels which never used a token.

5. Discussion

We applied NLP-based feature engineering to a dataset of IT certification items, fit a Naïve Bayesian Classifier (NBC) model, and cross-validated the model in a hold-out sample from the same domain and to a dataset of diverse items. The model performed reasonably well in cross-validation within the same domain (77.5%), although classification into Bloom's level 2 was weak (only 29% of level 2 items were classified correctly). A model designed to distinguish level 1 from higher levels (i.e., level 2 and above) was somewhat more effective (85.4% accuracy). But our IT certification NBC model generalized poorly to a new domain (overall accuracy was 21.3%). The feature engineering methods we used could be applied to any domain and do not require the preparation of a domain-specific knowledge base. Our results could be seen as an automated method for contextualizing Bloom's taxonomy for use in a specific domain and the Naïve Bayesian model was sufficiently interpretable to provide insights about the classifications.

We introduced this research as a way to improve the reliability of classifications into Bloom's taxonomy for exam items while reducing the time and labor requirements. The accuracy of about 78%, while being far less than 100%, must be sufficient for general use because this is comparable with the reliability of Bloom's classification performed by subject matter experts, although some programs may prefer that this is a second opinion that either confirms a SME judgment or causes the classification of an item to be more carefully reviewed. Either way, this approach shows great promise to improve both the efficiency and reliability of classification. One catch, however, is that this method can only be applied to a sizable pool that has been manually classified. We used a training sample of 100 items. Thus, the greatest efficiency gains may be had in programs where item-writing and Bloom's level classifications are an on-going issue, or where an existing pool of labeled items are available. On the other hand, having obtained adequate results with a sample of 100 items means this modeling is likely feasible for most exam programs, which often have larger pools.

Automated classification may also benefit programs where it is important to identify the Bloom's level of an item early. For example, if an exam blueprint specifies the Bloom's level of items written for an exam objective, then it would greatly improve the efficiency of item-writing if the level of the item were automatically checked as soon as the item was written. If an inappropriate level were detected, an automated warning could be given, and the SME author could then adjust the item before submitting it.

A theme in prior research on the reliability of Bloom's level classification has been to fold the original six levels into a streamlined taxonomy of three levels. Because of a lack of items in Bloom's levels 4-6 in our dataset, we collapsed these items into a single class. It seems very likely that this also improved the performance of our NBC model, and we would see reduced performance for a model predicting all six levels, particularly if there were also different base rates (e.g., most items were

written to address lower- or middle-levels). In practice, in the absence of a compelling rationale for using a six-level taxonomy, practitioners should consider a collapsed taxonomy with two to four levels. Our study and the prior literature provide a few choices:

- level 1; levels 2+3; levels 4+5+6 (Karpen & Welch, 2016)
- levels 1+2; levels 3+4; levels 5+6 (Plack *et al.*, 2007)
- level 1; levels 2+3+4+5+6 (current paper)
- level 1; level 2; level 3; levels 4+5+6 (current paper)

In our study, cross-validated accuracy rose to over 85% when we fit a dichotomous model classifying items into Bloom's level 1 vs. level 2 and above. Because this is a modest improvement over the 77.5% seen for the "full" model (which needed to collapse levels 4, 5, and 6), this result will mainly affect programs that are primarily concerned about identifying level 1 items.

We chose the Naïve Bayesian model because it was (relatively) interpretable. If we had chosen an artificial neural network, for example, it would have been nearly impossible to interpret the model. A decision tree is often touted as an interpretable model, but most real-world applications of decision or classification and decision trees are too complex to easily understand (see Sinharay 2016 for an example). This is partially because other models incorporate information about interactions that are ignored by "naïve" Bayesian predictive models.

A reviewer asked about the use of Bloom's taxonomy in writing these two samples. If Bloom's taxonomy was used, the version may affect our results. On the other hand, if the item writers did not target specific levels of Bloom's taxonomy then these results will especially generalize to pools written under similar conditions. We have little information about how these items were written. The items of the cross-domain dataset were heterogeneous and probably varied in the authors' use of Bloom's taxonomy.

Regarding the verbs associated with Bloom's levels, in this study verbs were important but not to the degree that they are emphasized in the literature on Bloom's taxonomy. The verbs associated with Bloom's levels are more likely to apply to learning objectives than items written to assess those objectives. For example, it is very plausible that a Level 3 learning objective about applying psychometric theory to reliability would use verbs like "apply" and "choose." But an item written to address this point is unlikely to use these verbs. For example, a scenario might specify circumstances and then the question might be "Which of the following is the best way to improve the reliability of these exam scores?" Instead, differences in word usage across levels seemed to be the main determinant of classification. In the examples presented in Tables 5 and 6, words absent from some levels exerted a strong influence by eliminating those levels.

One challenge and opportunity for psychometricians applying machine learning models is the need to "tune" ML models. For example, to avoid arithmetic underflow, most of the software packages (e.g., "e1071" in R) calculate conditional probabilities by adding logarithms of the probabilities (Manning *et al.*, 2009), which requires non-zero probabilities because log(0) is undefined in that case. In longer documents (with many tokens) a zero conditional probability can have an out-sized effect. And zero conditional probabilities may simply be due to not having a very large sample size. In practice, zero probabilities are not uncommon, and one way to deal with observed zero probabilities is to replace them with a very small value. Another solution is Laplace smoothing, where a constant "pseudocount" is added to each class in Equation (7). For example, if a smoothing constant of 1.0 is used, the smallest conditional probabilities will be 1/N. In trial runs (in this study, and other studies), we have observed that both practices slightly lowered crossvalidation performance. Formal or informal "tuning" of such parameters is a well-recognized aspect of ML models (Orrù *et al.*, 2020; Putka *et al.*, 2018). While parameter tuning is sometimes suggested as an advantage of machine learning models, tuning represents a challenge to practitioners. We strongly recommend that as psychometricians incorporate ML models into their research, that they also report on their tuning of model parameters.

5.1 Implications for Researchers and Practitioners

One open question about application of this technology to exam development is how a NBC model will change over the lifetime of an item pool, especially in the face of non-random changes to the items. For example, imagine that a domain should not include Bloom's level 1 items and the exam program implements a check based on such a model that informs item-writers when their item seems to be a level 1 item. Item writers are likely to

make the smallest possible change to the item to ensure that it measures comprehension or application. It seems far less likely that the item will be modified to measure cognitive processes of a much higher Bloom's level. This may influence the item pool in ways that cause the NBC model to become invalid over time. For example, we saw that "Which of the following…" items were more likely to be classified as level 1 because the tokens "follow" and "option" were disproportionately associated with level 1. If those terms are used in items that are then modified to be level 2 or higher, then the model's predictions will become less accurate over time. As a result, it may become important to continue to update the model as new items are added to the item pool.

This paper illustrates how a simple model, implemented in freely-available R packages, can be used to predict important aspects of exam items. Although the topic of this paper is the prediction of Bloom's taxonomy, the same methods can be used to predict other important aspects of exam items, such as difficulty, item quality, appropriateness of an item's topic, etc. These predictions could be made using the same methodology, although it is possible that different feature engineering steps might be required to improve the performance of these different models. For example, to predict the appropriateness of an item's topic, it seems likely that a corpus of written information (e.g., learning materials) classified by topic may be needed to effectively train a classifier with adequate accuracy and domain-specific stopword lists may be needed. However, there seems little doubt that predictions like these will play an increasingly important role in improving the exam development process.

5.2 Limitations and Future Directions

It is reasonable to assume that results will be affected by the specific item domain, the size of the item samples, and the distribution of items across Bloom's levels. Our item samples were on the smaller end of operational item pools. Therefore, these results need replication in additional samples of varying size to understand the range of outcomes that are likely and the effect of item sample size on the results. We are currently extending this research to encompass a greater diversity of exam items and explore the generalizability of the ML models to new exams within the same general domain and across domains. We are also examining additional feature engineering to support robust ML models that do generalize across domains.

We chose the Naïve Bayesian model because it is easily trained, relatively easy to interpret, and has a track record of success, but where interpretability is less important and computational resources are available, other models like support vector machines (SVM) and various types of artificial neural networks (such as convolutional neural networks; CNN) may produce higher levels of accuracy by using more parameters.

In addition, it is possible that more sophisticated feature engineering, especially in larger samples of items, might produce better results. In larger samples, considerable additional context can be embedded in the model by using n-grams as tokens. Bigrams and trigrams, for example, preserve the order of word pairs and triples. If most of the important interactions between words are present in word pairs or triples, the use of bigrams or trigrams will capture this source of information. As an example of this type of interaction, using unigrams, the phrases words "not safe" and "safe" are coded in a way that fails to capture the negation, but bigrams and trigrams could capture this negation. However, many bigrams and trigrams are likely to be unique (and thus not helpful) and larger samples of items would be needed to fully utilize this feature engineering approach.

6. References

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Abridged Edition). New York: Longman.

Athanassiou, N., McNett, J., & Harvey C. (2003). Critical thinking in the management classroom: Bloom's taxonomy as a learning tool. *Journal of Management Education, 27*(5), 555-575. https://doi.org/10.1177/1052562903252515

Billings, M. S., DeRuchie, K., Hussie, K., Kulesher, A., Merrell, J., Morales, A., Paniagua, M. A., Sherlock, J., Swygert, K.A., & Tyson, J. (2020). Constructing Written Test Questions for the Basic and Clinical Sciences (6th ed). Philadelphia, PA: National Board of Medical Examiners

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational

objectives; the classification of educational goals; Handbook I: Cognitive domain. New York, NY: Longmans, Green.

Buckwalter, J. A., Schumacher, R., Albright, J. P., & Cooper, R. R. (1981). Use of an educational taxonomy for evaluation of cognitive performance. *Journal of Medical Education, 56*(2), 115-21. https://doi.org/10.1097/00001888-198102000-00006

Chang, W., & Chung, M. (2009). Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items. *2009 Joint Conferences on Pervasive Computing (JCPC)*, 727-734. https://doi.org/10.1109/JCPC.2009.5420087

Cilliers, F. J., Schuwirth, L. W., Herman, N., Adendorff, H. J., & van der Vleuten, C. P. (2012). A model of the pre-assessment learning effects of summative assessment in medical education. *Advances in Health Sciences Education: Theory and Practice,* 17(1), 39-53. https://doi.org/10.1007/s10459-011-9292-5

Crowe, A., Dirks, C., & Wenderoth, M. (2008). Biology in bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE Life Sciences Education, 7*(4), 368-381. https://doi.org/10.1187/cbe.08-05-0024

Haris, S. S., & Omar, N. (2015). Bloom's taxonomy question categorization using rules and n-gram approach. *Journal of Theoretical and Applied Information Technology,* 7(3), 401-407.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. (2014). Teaching to the test… or testing to teach: Requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*, 307-329. https://doi.org/10.1007/s10648-013-9248-9

Jurafsky, D., & Martin, J. H. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice Hall.

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. *Currents in Pharmacy Teaching and Learning, 8*, 885-888. https://doi.org/10.1016/j.cptl.2016.08.003

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

Manning, C. D., Raghavan, P., & Schütze, H. (2009). An introduction to information retrieval (Online Edition). Cambridge University Press: Cambridge, England. https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Mesic, V., & Muratovic, H. (2011). Identifying predictors of physics item difficulty: A linear regression approach. *Physical Review Special Topics - Physics Education Research,* 7. https://doi.org/10.1103/PhysRevSTPER.7.010110

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLoS ONE. 15*(3). https://doi.org/10.1371/journal.pone.0230442

Nkanginieme, K. E. O. (1997). Clinical diagnosis as a dynamic cognitive process: Application of Bloom's Taxonomy for educational objectives in the cognitive domain. *Medical Education Online, 2*:1. https://doi.org/10.3402/meo.v2i.4288

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology, 10*, 2970. https://doi.org/10.3389/fpsyg.2019.02970

Plack, P.M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., & Greenberg, L. (2007). Assessing reflective writing on a pediatric clerkship by using a modified Bloom's Taxonomy. *Ambulatory Pediatrics, 7*, 285-291. https://doi.org/10.1016/j.ambp.2007.04.006

Porter, M.F. (1980). An algorithm for suffix stripping. *Program, 14*, 130-137. https://doi.org/10.1108/eb046814

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods, 21*, 689-732. https://doi.org/10.1177/1094428117697041

Rosca, C. V. (2004). *What makes a science item difficult? A study of TIMSS -R items using regression and the linear logistic test model.* Unpublished doctoral dissertation, Boston College.

Sinharay, S. (2016). An NCME Instructional Module on Data Mining Methods for Classification and Regression. *Educational Measurement: Issues and Practice, 35*, 38-54. https://doi.org/10.1111/emip.12115

Stephens, C. R., Huerta, H. F., & Linares, A. R. (2018). When is the Naïve Bayes approximation not so naïve? *Machine Learning,* 107, 397-441. https://doi.org/10.1007/s10994-017-5658-0

Tan, Y. T., & Othman, A. (2013). The relationship between complexity (taxonomy) and difficulty. *AIP Conference Proceedings, 1522*, 596. https://doi.org/10.1063/1.4801179

Thompson, E., Luxton-Reilly, A., Whalley, J., Hu, M. & Robbins, P. (2008). Bloom's taxonomy for CS assessment. Paper presented at Conference on Australasian computing education.

Yahya, A. A., Osman, A., Taleb, A., & Alattab, A. A. (2013). Analyzing the cognitive level of classroom questions using machine learning techniques. *Procedia – Social and Behavioral Sciences,* 97, 587-595. https://doi.org/10.1016/j.sbspro.2013.10.277.

Yahya, A. A., & Osman, A. (2011). Automatic classification of questions into Bloom's cognitive levels using support vector machines. *Proceedings of the International Arab Conference on Information Technology. Riyadh, Saudi Arabia*, 335-342. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.657.25&rep=rep1&type=pdf

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

Zaidi, N. L. B., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., Gruppen, L. D., & Santen, S. A. (2018). Pushing critical thinking skills with multiple-choice questions: Does Bloom's Taxonomy work? *Academic Medicine, 93*(6), 856-859. https://doi.org/10.1097/ACM.0000000000002087