An Accurate Classification and Recommendation Method of Competitive Math Problems

1st Yourui Shao
BASIS Independent Silicon Valley

San Jose, United States
youruishao115022@gmail.com

***Abstract*—Personalized and feedback-based learning frequently produces better learning outcomes in the classroom. Digital adaptive learning systems rely upon the categorization of problems and course materials to accurately recommend personalized content to users. While adaptive learning systems have already been widely adopted in math, the problems are often manually categorized and created. This paper discusses the multi-label classification application for categorizing competitive math problems into widely recognized categories using problem text, answer choices, and user-generated solutions with DistilBERT, a Transformer-based language representation model. Additionally, by adding LaTeX control sequences to the tokenizer vocabulary, the trained model was able to recognize symbolic math equations and thus more accurately predicted problem categorization compared to using a tokenizer without additional vocabulary. The trained model with additional vocabulary mostly accurately predicted problem labels across all five general categories, thus demonstrating DistilBERT's ability for problem categorization and the significance of representing symbolic equations with vocabulary. Based on the classification results, a problem recommendation algorithm adaptive to user performance is introduced, to illustrate an application of categorization in adaptive learning systems.**

***Index Terms*—math problem classification, adaptive learning platforms, DistilBERT**

I. Introduction

Education in mathematics, among other subjects, has relied on the repetition of practice problems to increase students' subject familiarity. Performance-based feedback and personalization have been shown to both improve student outcomes. In a combination of earlier studies, a "two-sigma" problem emerged, with students learning from a personal tutor performing two standard deviations above the control mean on average compared to lecture-based and revision-based classroom settings [1]. Now, intelligent tutoring systems have been developed to replicate the human tutor, supplying educational materials with characteristics based on the parameters supplied by student performance, motivation, and behaviors. [2].

A subject that popularly leveraged intelligent tutoring systems—or adaptive learning platforms—is mathematics. Khan Academy, an online repository of educational videos and over 100,000 math practice problems uses immediate problem feedback, hints, and referrals to relevant videos when practicing [3]. In one study, 85% of teachers surveyed reported the integration of Khan Academy into their classrooms for reinforcing math topics had a positive effect on student understanding and learning [3]. Alcumus, developed by the Art of Problem Solving, is a digital math practice platform that maps problems to math topics to provide students with topically related problems and adapt to student skill levels using machine learning [4].

By recommending problems based on student performance, these platforms customize learning to users on a large scale. The mastery system in Khan Academy allows students to set the learning pace to match each student's level of understanding [3]. Similarly, Alcumus calculates a score corresponding to multiple levels of mastery for each individual math topic. Within problems belonging to a topic, those with a difficulty appropriate for a student's score in the topic are recommended based on logistic regression of the likelihood that the student answers correctly [4].

In Khan Academy and Alcumus, the topics are grouped into topic groups and subjects, respectively. As such, classification and grouping of problems into different subjects or topics is critical to both adaptive learning platforms.

High school math competitions typically feature problems testing the depth of knowledge and high-level problem-solving skills [5]. Competitive math is typically segmented into distinct categories, subjects which in themselves hold a variety of difficulties. The Mathematical Association of America, the organization that creates and hosts the AMC 10, states that the AMC 10 assumes the knowledge of elementary algebra, basic geometry, elementary number theory, and elementary probability [6]. The problems in textbook *Competition Math For Middle School* are based on the "AMC series," in which the AMC 10 is included. It is segmented into the five chapters of "Algebra," "Counting," "Probability," "Number Theory," and "Geometry," similar to the topics tested on the AMC 10 as stated by the MAA [7]. As will be discussed in Hierarchical Categorization, this paper will use general categories based on Batterson's chapters.

This paper proposes a context-based method of effectively categorizing math problems and a possible system for recommendation. We will first discuss the classification of math problems in other works using other problem sets and set categories. Then, we will explain the collection and cleaning process for AMC 10 problems as well as the treatment of symbolic equations with additional vocabulary during tokenization; the DistilBERT model and its suitability for language

classification tasks; and the training process and performance metrics of the experimental multi-label classification models. Additionally, we describe one method in which, using categories, math problems can be recommended to users in an adaptive learning system. Lastly, we propose methods for future improvement and application to adaptive learning systems in education.

II. Related Work

Previously, various methods have been used to categorize elementary math problems into a small number of categories based on the solving method. For the development of intelligent math tutors, Cetintas et al. trained a Support Vector Machine (SVM) model to classify "multiplicative compare" and "equal group" elementary word problems and stemming and eliminating parts of speech non-discriminatory between problem types [8]. By combining multiple SVM techniques, an average F_1 score of .795 was achieved.

Nandi and Narang used and compared SVM, Decision Tree, k-Nearest Neighbour, Neural Network, and Convolutional Neural Network models to categorize word problems by the method of solution—"addition," "subtraction," "multiplication," and "division"—not explicitly stated by the problem [9]. Problems were represented as vectors produced by the bag-of-words technique, which outputs the frequency of words in a constant word list within the given problem. Non-discriminatory parts of speech were eliminated as was done in Cetintas et al., but the procedure was modified to include newly discriminatory keywords for this particular classification task, such as "average" and "each," which were not useful for the "multiplicative compare" and "equal group" tasks. For these tasks, the simple neural network achieved a low misclassification error of 0.0050 [9].

More recently, by isolating relevant terms via the Term Frequency—Inverse Document Frequency method, then building a Word Mover's Distance on the isolated terms to vectorize math problem text, Costa et. al applied various classification models to label elementary school math problems with the Computational Thinking Skills involved in completing the problem [10]. While the XGBoosting model obtained the highest average precision of 92.46% overall, it also featured the lowest precision score for individual categories, with 72.50% precision for the "Decomposition" skill.

This paper focuses on the classification of problems within larger, general math subjects typically covered in secondary education rather than specific arithmetic tasks in early education. Additionally, by utilizing DistilBERT, an attention-based model, context-heavy distinctions can be made to categorize problems more accurately.

Chen et. al discusses the usage of a collaborative filtering algorithm in an adaptive learning system. This algorithm creates a student-score matrix that indicates whether a student is interested in a subject. The "neighboring" students, who share the most similar student-score interest matrix, are identified, and a K-means algorithm was used to optimize the collaborative filtering of users and items. [11]

III. Methods

A. Data Preparation

1) Data Source: **The AMC 10 Exam** The AMC exams are composed of 25 questions, which are ordered from 1 through 25 and are designed to increase in difficulty. Each question is answered by selecting one of five provided choices, numbered `A` to `E`. Since 2002, two AMC 10 exams have been administered a year with different exam problems except in 2021, in which four exams were administered in total.

Sourcing AoPS.com hosts an AoPSWiki, to which account-holding AoPS.com users submit questions recollected after the exams are administered [12]. Each problem on each exam is traditionally assigned a page with a unique web URL. A problem's page contains the exam information, the problem statement, the five choices available for selection, and a varying number of solutions that account-holding users compile and upload [12] (Figure1).

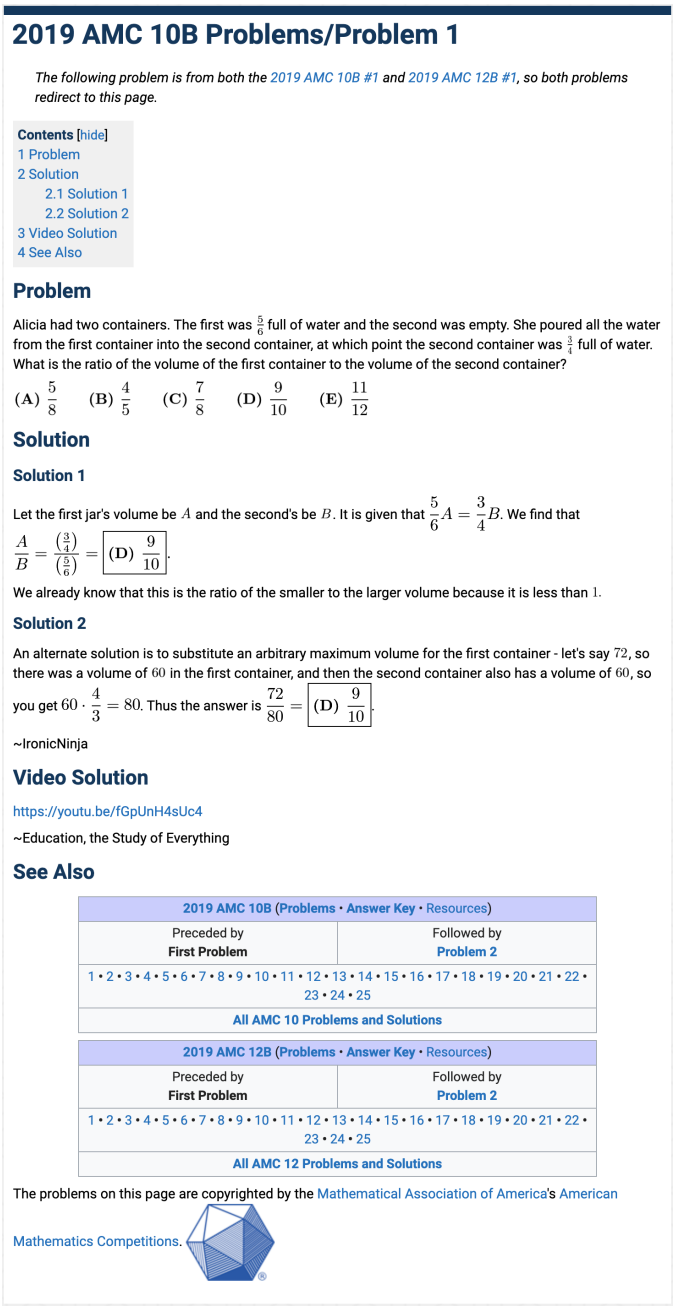
LaTeX Equations Symbols in mathematics are commonly notated using LaTeX or other TeX-based typesetting software [13]. In LaTeX, control sequences, such as `\frac{}{}`, often beginning with a backslash (`\`) with exceptions such as `^` and `_` are used to typeset standard symbols and character positioning used mathematics. For example, the equation $c^2 = a^2 + b^2 - 2ab\cos C$ is rendered with the LaTeX syntax of `c^2 = a^2 + b^2 -2ab\cos C`, and a natural language representation could be "c squared equals a squared plus b squared minus two times a times b times the cosine of C."

Math problems frequently contain symbolically represented equations. For problems in particular categories, the content of equations may provide clues to its categorization. For example, the use of $\cos$, representing a cosine function, in a problem frequently indicates the involvement of geometry. As such, retaining the equations is assistive to problem classification.

As the data source for problems on the AMC 10 commonly renders equations via LaTeX, as is done when the contests are administered, a plain text representation of equations can be achieved by translating equations to natural language or by including the raw LaTeX syntax directly. As little LaTeX-to-natural language technology exists for this purpose, and for the simplicity of the process, in this paper, the raw LaTeX syntax was included directly. The DistilBERT pre-trained tokenizer does not recognize LaTeX control sequences in input text natively and rather splits these tokens (Figure 4). As such, the vocabulary of the tokenizer was expanded so LaTeX control sequences could be properly tokenized and recognized in the training process (Section III-A3).

Data Collection Through accessing the HTML components of the page, different parts of the problem page—the problem statement, choices, and solutions—can are isolated.

As problem pages are edited and formatted by different account-holding users across multiple years, some issues arise with inconsistent syntax. Problems in exams administered before 2015 were represented more differently syntactically which made collection from HTML difficult. As such, only problems from exams between 2015 A and 2022 B, inclusive,



2019 AMC 10B Problems/Problem 1

The following problem is from both the 2019 AMC 10B #1 and 2019 AMC 12B #1, so both problems redirect to this page.

Contents [hide]
1 Problem
2 Solution
2.1 Solution 1
2.2 Solution 2
3 Video Solution
4 See Also

Problem

Alicia had two containers. The first was $\frac{5}{6}$ full of water and the second was empty. She poured all the water from the first container into the second container, at which point the second container was $\frac{3}{4}$ full of water. What is the ratio of the volume of the first container to the volume of the second container?

$\textbf{(A)}\ \frac{5}{8} \quad \textbf{(B)}\ \frac{4}{5} \quad \textbf{(C)}\ \frac{7}{8} \quad \textbf{(D)}\ \frac{9}{10} \quad \textbf{(E)}\ \frac{11}{12}$

Solution

Solution 1

Let the first jar's volume be A and the second's be B. It is given that $\frac{5}{6}A = \frac{3}{4}B$. We find that $\frac{A}{B} = \frac{\left(\frac{3}{4}\right)}{\left(\frac{5}{6}\right)} = \boxed{\textbf{(D)}\ \frac{9}{10}}$.

We already know that this is the ratio of the smaller to the larger volume because it is less than 1.

Solution 2

An alternate solution is to substitute an arbitrary maximum volume for the first container - let's say 72, so there was a volume of 60 in the first container, and then the second container also has a volume of 60, so you get $60 \cdot \frac{4}{3} = 80$. Thus the answer is $\frac{72}{80} = \boxed{\textbf{(D)}\ \frac{9}{10}}$.

~IronicNinja

Video Solution

https://youtu.be/fGpUnH4sUc4

~Education, the Study of Everything

See Also

2019 AMC 10B (Problems • Answer Key • Resources)	
Preceded by First Problem	Followed by Problem 2
1 • 2 • 3 • 4 • 5 • 6 • 7 • 8 • 9 • 10 • 11 • 12 • 13 • 14 • 15 • 16 • 17 • 18 • 19 • 20 • 21 • 22 • 23 • 24 • 25	
All AMC 10 Problems and Solutions	

2019 AMC 12B (Problems • Answer Key • Resources)	
Preceded by First Problem	Followed by Problem 2
1 • 2 • 3 • 4 • 5 • 6 • 7 • 8 • 9 • 10 • 11 • 12 • 13 • 14 • 15 • 16 • 17 • 18 • 19 • 20 • 21 • 22 • 23 • 24 • 25	
All AMC 12 Problems and Solutions	

Figure 1. The problem page for number 1 on the 2019 B exam on AoPS.com.

were collected. Additionally, some wiki pages were manually modified to adhere to a consistent syntactical standard.

Solutions Often only contain a link to an external video without additional context. These solutions are removed for the data used in model training. For the 449 problems collected, there were 2602 solutions in total, of which 1468 were removed due to containing these external links.

2) Data Transformation: **HTML** Tags In preparation for tokenization, any HTML tags attached to problems or their solutions were removed. LaTeX equations are rendered through `<img>` HTML tags with the raw LaTeX included in the "alt" attribute, and the syntax is isolated and inserted into the rest of the problem at the location of the equation (Figures 2 and 3).

```
<img alt="$\tfrac{5}{6}$" class="latex"
height="21" src="//latex.artofproblemso
lving.com/9/7/0/970ffc67a479e2152f64973
7fb6fa0acc4bb961f.png" style="vertical
-align:-6px" width="9"/>
```

Figure 2. Example of use of `<img>` tag to represent LaTeX, from problem number 1 on the 2019 B exam

```
<html><head></head><body><p>Alicia had
two containers. The first was <img
alt="$\tfrac{5}{6}$" class="latex"
height="21" src="//latex.artofproblem
solving.com/9/7/0/970ffc67a479e2152f6
49737fb6fa0acc4bb961f.png" style="ver
tical-align: -6px" width="9"/> full
of water ... </p><p> </p></body></html>
```

Problem statement as unprocessed HTML

```
<html><head></head><body><p>Alicia had
two containers. The first was \tfrac
{5}{6} full of water ... </p><p> </p>
</body> </html>
```

Problem statement with `<img>` tag converted into only LaTeX syntax

```
Alicia had two containers. The first
was \tfrac{5}{6} full of water ...
```

Problem statement without any HTML tags

Figure 3. Conversion of `<img>` tag representing LaTeX and removal of HTML tags in preparation for tokenization, with problem number 1 on the 2019 B exam.

Total Context The context of a problem combines the three text-based segments of a problem—the problem statement, the user-generated solutions which are not solely composed of external links, and the five answer choices. Stopwords from the NLTK Python package, which features 179 general stopwords such as "a, an, the, ...what, which who...," was removed from the context.

3) Tokenization: The pre-trained, WordPiece-based DistilBERT tokenizer was modified and used to tokenize the context.

Vocabulary LaTeX control sequences may assist in interpreting and identifying a problem type, as discussed in data source part. Without additional vocabulary, the tokenizer frequently splits up LaTeX control sequences into multiple tokens. As such, a list of 13622 LaTeX symbols from CTAN and 203 amsmath supplemental control sequences were added

to the tokenizer's vocabulary [14] [15]. Figure 4 compares the tokenized problem context with and without the addition of control sequences to the vocabulary, for number 1 on the 2019 B exam.

```
"Alicia two containers. first \\tfrac
{5}{6} ..."
```

Raw context

```
['alicia', 'two', 'containers', '.',
'first', '\\', 't', '##fra', '##c',
'{', '5', '}', '{', '6', '}', ...]
```

DistilBERT tokenizer output without added vocabulary

```
['alicia', 'two', 'containers', '.',
'first', '\\tfrac', '{', '5', '}', '{',
'6', '}', ...]
```

DistilBERT tokenizer output with additional vocabulary

Figure 4. Tokenization outputs for the context of problem number 1 on the 2019 B exam.

4) Data Labeling: Four general categories (GCs) of AMC 10 problems—"Algebra," "Geometry," "Counting and Probability," and "Number Theory"—were adapted from classifications (see Section I). Problems within a GC test separate "spheres of knowledge." We can identify sub-categories (SC) within each GC that pertain either to specific solution patterns or problem features. A fifth GC, "Miscellaneous," was added to encompass problems identified with SCs such as "Arithmetic," "Word Problems," and "Logic," not typically recognized as belonging to any of the four recognized GCs. The labeling information can be found in Table I. By introducing hierarchy, a classification model can first categorize a problem with its GCs, and then SCs within the identified GCs.

GC	Frequency	SCs
Miscellaneous	97	Arithmetic, Word Problems, Fractions/Percentages, Tracing, Strategy/Answer Choices, Exponentiation/nth Roots, Logic, Sequencies and Series, Geometric Sequence, Arithmetic Sequence, Telescoping Series, "Mean, Median, Mode, Range", Inequalities, Roots of Unity, Binomials, Complex Numbers, Symmetry, Estimation, Pattern Finding
Algebra	62	Arithmetic, Word Problems, Fractions/Percentages, Tracing, Strategy/Answer Choices, Exponentiation/nth Roots, Logic, Sequencies and Series, Geometric Sequence, Arithmetic Sequence, Telescoping Series, "Mean, Median, Mode, Range", Inequalities, Roots of Unity, Binomials, Complex Numbers, Symmetry, Estimation, Pattern Finding
Geometry	66	Angle Chasing, Similar/Congruent Triangles, Regular/Equiangular Polygons, Triangles – Special Points and Lines, 3D Geometry, Regular Polyhedrons, Trigonometry, Trigonometric Identies, Analytical Geometry, Arcs, Circles, Inscribed and Circumscribed Circles, Pythagorean Theorem, "Reflections, Translations, and Rotations", Area, Polar Coordinates
Number Theory	39	Modular Arithmetic, Divisibility, Digits, Primes, Divisors, Base Conversion, GCDs and LCMs
Counting and Probability	71	Counting*, Probability*, Casework, Configuration/Symmetry (Burnside's), Path Counting, Dice Problems, Card Problems, Placing/Picking/Labeling Problems, Stars and Bars, Expected Value, Recursive Counting, Conditional Probability, Pigeonhole Principle, Complementary Counting, States, Principle of Inclusion-Exclusion, Geometric Probability

*Although not technically SCs of "Counting and Probability," the categories "Counting" and "Probability" are used to distinguish math problems that belong to only one of the two.

Table I
THE FIVE GCS, THE FREQUENCY OF PROBLEMS THAT BELONG TO THE GC IN THE LABELED DATASET, AND THE SCS OF THAT GC.

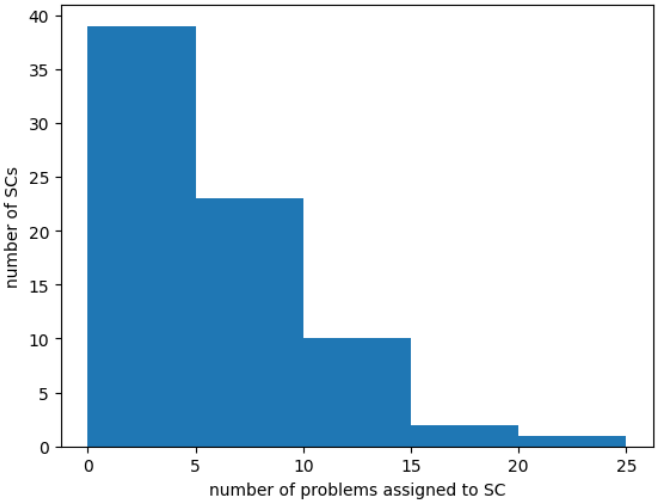


Figure 5. The frequency of SCs from the 250 labeled problems.

The 250 problems from exams between the 2019 A and 2022 B were labeled manually and used in training the model. For each SC listed, each problem was identified as belonging to or not belonging to the SC. For a problem to be labeled by an SC, it must also be labeled by the GC to which the SC belongs. For example, for a problem to be labeled with "Polynomial," it must also be labeled as "Algebra." For a list of all SCs, see Table I. Each problem may also be labeled with multiple SCs within a general category, or multiple GCs. However, as SCs were often assigned to a small number of problems during labeling, with only 11 SCs assigned to 10 or more problems, the low positive sample size makes SC classification impractical for most SCs with the current dataset (Figure 5).

B. Model Selection and Training

The Transformer architecture has allowed language representation models to be pretrained on generic corpora and then

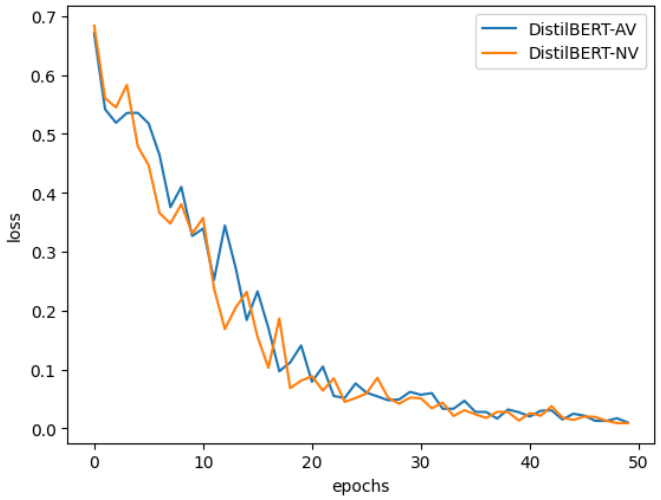


Figure 6. Loss from epoch 1 to 50 while finetuning DistilBERT on the GCs with an 80-20 train-test split, for the NV and AV models.

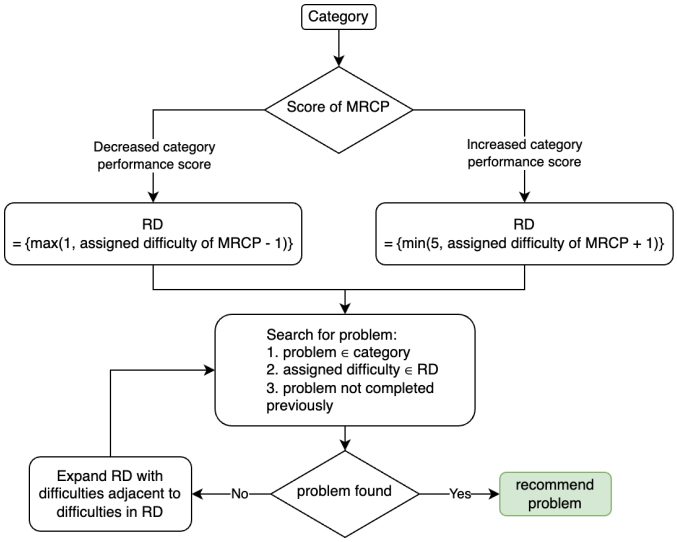


Figure 7. Process for selecting a problem within a given category, with MRCP as the most recently completed problem in the category, and RD as the list of possible difficulties for the problem to recommend.

easily fine-tuned downstream for a variety of tasks [16]. One such model is BERT, which uses mask language modeling, a technique that teaches the model to predict tokens masked at random based on the surrounding context [17]. This allowed the model to recognize bidirectional context at a given token rather than reading the context right-to-left and left-to-right. As such, it can be fine-tuned with a single additional output layer to perform a variety of language tasks, such as question answering and text classification.

DistilBERT was introduced in 2019 as a smaller, faster distillation of the BERT transformers model, which was used as a teacher for DistilBERT's self-supervised pretraining on the BookCorpus dataset [18]. Like BERT, DistilBERT can be used for English-language classification tasks without complex modifications due to the bi-directional pretraining and was thus used in this paper. From the distillation process, DistilBERT is also smaller and faster at making inferences. Unlike previous keyword-tagging methods for classifying math problems discussed in Related Work, DistilBERT can contextualize words in relation to each other and thus serves complex tasks in which the position of words is significant. The DistilBERT uncased model was trained to classify GCs in the context of the 250 labeled problems with an 80-20 train-test split in 50 epochs (Figure 6). Two models were trained, one with added LaTeX control sequences in the vocabulary (AV) and one without the additional vocabulary (NV). The model outputs the true and false states of each category for the given text.

C. Problem Recommendation

The purpose of problem classification is to automatically recommend suitable practice problems to user. In this use case, we design and implement a practice problem recommendation system. Thus these classified problems were incorporated into the recommendation system that continuously supplies users with a problem. In summary, the recommended problem is based on their category performance scores, calculated after the previous problem is completed which considers the problem difficulty level.

1) Performance Scores: All problems are assigned a measure of difficulty, which is estimated using the number of the problem within a given exam, as problems are generally ordered in increasing difficulty as determined by exam writers. Within the 25-problem exam, the first five problems receive a difficulty score of 1; problems six through ten, inclusive, receive a difficulty score of 2, and so on; with the last five problems receiving a difficulty score of 5.

After the completion of a problem, a problem performance score is calculated based on the problem's preassigned difficulty and the number of attempts taken for the user to answer correctly based on Equation 1. There are five answer choices for a problem, and users here have been prevented from re-entering an incorrect guess. We maximize the possible base score at 100, where the user answers correctly with one attempt, and deduct 25 points for each additional attempt taken. Then, we apply based on preassigned problem difficulty a multiplicative bonus, which adds 25% of the base score for each point increase in difficulty above 1. Thus, the maximum multiplied score for a problem is 200, and the lowest score is 0.

$$25 \cdot (5 - \text{attempts})(1 + .25 \cdot (\text{problem difficulty} - 1)) \quad (1)$$

The category performance is the mean problem performance in all problems completed in that category.

2) Problem Selection: To select the problem to recommend to the user, the category to which the next problem belongs is selected, from which then a problem is chosen according to the appropriate difficulty. Each category is assigned a weight equal to its performance score, and a category is randomly selected according to the assigned weights. Within the category selected, a problem that the user has not completed prior and

Category		Accuracy	Precision	Recall
Overall	NV	.94	.969	.827
	AV	.972*	.986*	.92*
Miscellaneous	NV	.94	1	.893
	AV	.98*	1	.964*
Algebra	NV	.92	1	.667
	AV	.94*	.909†	.833*
Geometry	NV	.98	1	.917
	AV	.98	1	.917
Number Theory	NV	.96	.778	1
	AV	.98*	1*	.857†
Counting and Probability	NV	.90	1	.688
	AV	.98*	1	.938*

*Category increased in performance in the AV model from the NV model
†Category decreased in performance in the AV model from the NV model

Table II
THE ACCURACY, PRECISION, AND RECALL FOR THE PREDICTIONS OF THE AV AND NV MODELS ON THE 50 VALIDATION PROBLEMS OVERALL AND IN EACH GC.

in the assigned difficulty range determined by the process illustrated in Figure 7, is selected as the recommended problem. If the most recently completed problem in the category (MCRP) lowered the category performance, then a problem with assigned difficulty lower than that of the MRCP is added to the recommended difficulties (RD); and vice versa for an MRCP that raised the category performance. The RD is expanded until a viable problem is found. If no problem has been completed in the category, one of difficulty 3 is chosen at random.

IV. Results

The model was validated on 50 problems across all five GCs, thus outputting 250 labels in total. There is a significant number of false negatives produced by both the NV and AV models, while only one to two false positives arose out of all labels, indicating a generally high precision. Additionally, the number of false negatives decreased and the recall increased from the NV to the AV model from .83 to .92, indicating that the addition of LaTeX control sequences positively assisted with categorization.

Across individual GCs, the recall also increased when moving from NV to AV, except for Geometry, which retained the same recall rate, and Number Theory, which decreased in recall rate with one additional false negative. Overall, the accuracy of the NV model is .94 and .972 for the AV model, indicating a strong ability for the trained DistilBERT model with added vocabulary to accurately classify problems into one or more of the five GC categories. See Table II for a comprehensive calculation of the performance metrics.

V. Conclusion

In this paper, the DistilBERT model was trained to classify high school competition math problems with the AMC 10 exam. Multiple parts of a problem providing context to the category to which belongs: the problem statement, the choices, and the solutions, were concatenated as the inputs to be tokenized and used for classification. To increase recognition of math equations represented in LaTeX during classification, additional control sequences were added to the pre-trained DistilBERT tokenizer, which improved the classification recall. This demonstrates the benefit of including LaTeX equations during language processing of problems on math contests or other texts dependent on symbolic representation.

While the GCs of "Miscellaneous," "Algebra," "Geometry," "Number Theory," and "Counting and Probability" had problem frequencies sufficient for training the model and producing effective results as shown by the high recall and accuracy rates, SCs were not sufficiently frequent for individual models. A reevaluation of how SCs are categorized, or a larger labeled dataset may be needed to use DistilBERT to hierarchically categorize SCs once the GCs were identified. By consolidating multiple SCs or retaining only those that occur frequently, a larger number of positives would likely enable the model to more successfully detect patterns and thus categorize SCs. Similarly, expanding the labeled dataset could also allow successful categorization of SCs. Through the inclusion of SCs, the proposed recommendation system will also be able to more precisely identify categories to recommend to users.

Currently, the recommendation system prioritizes providing problems from a mix of categories based on a multiplicative weighting of category performance scores. Through implementing a feedback system in which users can indicate whether they believe a recommendation to be appropriate, as well as monitoring score improvement, adjustments could be made to the weighting of categories and the difficulty progression.

Problems from other contests such, as the AMC 12, could also be categorized and recommended using a similar approach. Problems of new, previously unrecognized categories can be easily incorporated to be identified parallel to the current GCs by retuning the DistilBERT model with a labeled dataset with sufficient problems belonging to the new category. While this paper focuses on applications of text classification to competitive math, similar approaches can be used to categorize problems and course materials in school math and other subjects. Categorization is important for the element of progression and recommendation in adaptive learning systems (see Section I), which recreate the benefits of personalized tutoring to improve student learning outcomes.

References

[1] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educational Researcher*, vol. 13, no. 6, pp. 4–16, 1984. [Online]. Available: http://www.jstor.org/stable/1175554

[2] S. Cetintas, "Effective and efficient user and content modeling for intelligent tutoring systems," Ph.D. dissertation, Purdue University, 2012.

[3] R. Murphy, L. Gallagher, A. E. Krumm, J. Mislevy, and A. Hafter, "Research on the use of khan academy in schools: Research brief," *Menlo Park, CA., SRI International. Recuperado de https://www. sri. com/sites/default/files/publications/2014-03-07_implementation_briefing. pdf*, 2014.

[4] Art of Problem Solving, "What is alcumus, and why we called it that." [Online]. Available: https://artofproblemsolving.com/blog/articles/what-is-alcumus-why-we-called-it-that

[5] G. Ellison and A. Swanson, "Do schools matter for high math achievement? evidence from the american mathematics competitions," *American Economic Review*, vol. 106, no. 6, pp. 1244–1277, 2016.

[6] "Amc 10/12." [Online]. Available: https://maa.org/math-competitions/amc-1012

[7] J. Batterson, *Competition math for middle school*. Alpine, CA: Art of Problem Solving, 2011.

[8] S. Cetintas, L. Si, Y. P. Xin, D. Zhang, and J. Y. Park, "Automatic text categorization of mathematical word problems," in *Twenty-Second International FLAIRS Conference*, 2009.

[9] B. P. Nandi and P. A. Narang, "Mathematical word problem categorization using machine learning," 2017.

[10] E. J. F. Costa, C. E. C. Campelo, and L. M. R. S. Campos, "Automatic classification of computational thinking skills in elementary school math questions," in *2019 IEEE Frontiers in Education Conference (FIE)*, 2019, pp. 1–9.

[11] W. Chen, Z. Shen, Y. Pan, K. Tan, and C. Wang, "Applying machine learning algorithm to optimize personalized education recommendation system," *Journal of Theory and Practice of Engineering Science*, vol. 4, no. 01, p. 101–108, Feb. 2024. [Online]. Available: https://centuryscipub.com/index.php/jtpes/article/view/464

[12] "Editing the aopswiki." [Online]. Available: https://artofproblemsolving.com/wiki/index.php/AoPS_Wiki:TutorialEditing_the_AoPSWiki

[13] H. Kopka and P. W. Daly, *Guide to LATEX*. Pearson Education, 2003.

[14] S. Pakin, "The comprehensive latex symbol list – symbols accessible from latex," Jan 2023. [Online]. Available: https://www.ctan.org/tex-archive/info/symbols/comprehensive

[15] LaTeX Project and American Mathematical Society, Dec 1999. [Online]. Available: http://www.ams.org/arc/tex/amsmath/amsldoc.pdf

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.