# Exploring Machine Learning Techniques for the Classification of Mathematical Texts

**Adrian Martushev**
**University of Oregon**
**amartush@uoregon.edu**

## Abstract

This project focuses on classifying mathematical problems by type using machine learning techniques. Utilizing a dataset comprising math problems categorized by algebra, geometry, and other subjects, the project employs various classification algorithms to predict the type of math problem based on its textual description. The dataset, compiled from a structured directory of JSON files, includes attributes like problem statement, difficulty level, type, and solution. Preliminary results indicate significant potential for machine learning models to automate the classification of educational content, enhancing educational resources and learning platforms' efficiency.

## 1 Introduction

The categorization of mathematical problems by type (e.g., algebra, calculus) is a significant first step toward developing models capable of generating accurate solutions. Furthermore, efficient classification of problems can lead to more personalized learning experiences and better resource allocation in educational settings. This project aims to apply natural language processing (NLP) and machine learning techniques to classify math problems based on their descriptions automatically. This study aims to explore how different machine learning models perform in understanding and categorizing textual data within educational content and provide a basis for creating models capable of solving these problems.

## 2 Related Work

Initial attempts at classifying mathematical problems primarily utilized rule-based systems, which employed hand-crafted features and heuristic algorithms to identify mathematical terms and classify problems based on these patterns. These early methods, while foundational, often lacked the flexibility and scalability provided by more modern machine learning approaches.

### 2.1 Transition to Machine Learning Models

Recent advancements have shifted towards using sophisticated machine learning models, particularly those leveraging natural language processing capabilities. Studies have explored various frameworks for understanding and solving math word problems (MWPs), which involve analyzing text descriptions and generating corresponding mathematical equations. A notable study by Zhou et al. focused on addressing the limitations of single-solver approaches in MWPs, which often struggled with the diversity of problem types and were prone to overfitting. Their work proposed an innovative ensemble approach that combines the strengths of tree-based solvers and large language models (LLMs) to enhance problem-solving capabilities across a broader spectrum of MWPs [Zhou et al., 2023][2].

# 3 Methods

## 3.1 Data Description and Preprocessing

The data for this project was sourced from the Mathematics Aptitude Test of Heuristics (MATH) dataset[1]. This dataset is a comprehensive collection of problems from mathematics competitions such as the AMC 10, AMC 12, and AIME, among others. It is specifically designed for training and evaluating machine learning models on mathematical problem-solving and reasoning. Each problem is accompanied by a detailed step-by-step solution, presented in both LaTeX and natural language, making it an ideal resource for tasks that involve generating mathematical derivations and explanations.

## 3.2 Data Splits

The dataset is divided into two main splits:

**Train**: Consisting of 7,500 examples, this set is used to train the machine learning models.

**Test**: Comprising 5,000 examples, this set is utilized to evaluate the models' performance and generalizability.

## 3.3 Data Preprocessing

The initial phase of data preparation included consolidating JSON files into two large files, one for each data split, to simplify access and manipulation during model training and testing. Subsequent preprocessing steps involved text normalization, where all characters were converted to lowercase and non-standard or extraneous characters were removed to ensure consistency across the data. The process also included tokenization, breaking down problem descriptions and solutions into tokens, which are the basic units for model input. Finally, vectorization was employed using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to transform these textual tokens into a numerical format that machine learning algorithms can effectively process.

# 4 Experiments

## 4.1 Baseline Model Evaluation

The baseline model for this project is a Logistic Regression classifier, chosen for its simplicity and effectiveness in many text classification tasks. This model was implemented to classify mathematical problems into one of seven categories: Algebra, Counting & Probability, Geometry, Intermediate Algebra, Number Theory, Prealgebra, and Precalculus. The Logistic Regression model was trained using the TfidfVectorizer from scikit-learn, which converts text data into TF-IDF features. The vectorizer was configured to ignore English stopwords and limit the feature set to the top 5000 terms, balancing computational efficiency with feature coverage. The logistic regression model was tuned with a maximum iteration parameter set to 1000 to ensure convergence. The regularization strength was set to the default value (C=1.0)

**Results**

**Overall Accuracy**: 69.56%

The results indicate that the model performs best in categories with well-defined mathematical terminology such as Precalculus, which achieved the highest F1-score. Categories with more varied language and less distinct terms, such as pre-algebra, posed more significant challenges, resulting in lower precision and recall. The following classification report details the baseline precision, recall, and f-1 score for the categories

| **Category** | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| Algebra | 0.65 | 0.75 | 0.70 | 1187 |
| Counting & Probability | 0.76 | 0.70 | 0.73 | 474 |
| Geometry | 0.66 | 0.75 | 0.70 | 479 |
| Intermediate Algebra | 0.75 | 0.73 | 0.74 | 903 |
| Number Theory | 0.73 | 0.75 | 0.74 | 540 |
| Prealgebra | 0.54 | 0.48 | 0.51 | 871 |
| Precalculus | 0.93 | 0.76 | 0.84 | 546 |
| **Accuracy** | - | - | **0.70** | **5000** |

### 4.2 BERT-Based Text Classification

**Model Description and Setup**

In an extension of the initial experiments with logistic regression, a more sophisticated model using BERT (Bidirectional Encoder Representations from Transformers) was incorporated, specifically the bert-base-uncased variant. This model is highly regarded for its ability to understand the nuances of language by pre-training on a large corpus and fine-tuning on specific tasks like text classification.

The BERT model was integrated using Hugging Face's Transformers library, which facilitated a streamlined approach for model training and evaluation. The training involved three epochs with a batch size of eight, optimized for memory efficiency and allowing the model to learn from a diverse set of examples within each training step. Additionally, learning rate adjustments were made by implementing 500 warmup steps to gradually increase the learning rate, thereby mitigating the potential for disruptive updates early in the training phase. Furthermore, a weight decay of 0.01 was applied as a form of regularization to reduce the risk of overfitting by penalizing larger weights.

Model performance was evaluated using the 5,000 test examples, ensuring that assessments of model efficacy were based on data not seen during the training phase. Evaluation was structured to occur at the end of each epoch to track and monitor the model's progress and make necessary adjustments.

**Results**

**Overall Accuracy**: 77%

The BERT-based model demonstrated a notable improvement in accuracy and class-wise metrics over the baseline logistic regression model in all categories:

| **Category** | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| Algebra | 0.78 | 0.75 | 0.76 | 1187 |
| Counting & Probability | 0.78 | 0.75 | 0.76 | 474 |
| Geometry | 0.73 | 0.86 | 0.79 | 479 |
| Intermediate Algebra | 0.86 | 0.81 | 0.83 | 903 |
| Number Theory | 0.65 | 0.91 | 0.76 | 540 |
| Prealgebra | 0.65 | 0.57 | 0.60 | 871 |
| Precalculus | 0.97 | 0.85 | 0.90 | 546 |
| **Accuracy** | - | - | **0.77** | **5000** |

**Discussion**

The application of BERT in classifying mathematical problems brought forward several insights. First, categories such as Number Theory, which often involve complex problem statements, saw high recall rates, indicating the model's strength in identifying relevant cases. Second, Prealgebra continued to present challenges, likely due to its broad and fundamental nature, which might not have distinctly unique linguistic features as compared to more advanced topics.

The experiment with BERT not only validated the hypothesis that transformer-based models could outperform more traditional machine learning approaches in text classification tasks but also highlighted the potential for further refinement in handling categories with overlapping or ambiguous terms. Future work will focus on expanding the dataset, experimenting with hyperparameter tuning, and exploring ensemble methods to combine the strengths of different model architectures.

## 5 Conclusion

This study explored the application of machine learning techniques for classifying mathematical problems into specific categories using the Mathematics Aptitude Test of Heuristics (MATH) dataset. The investigation started with a baseline logistic regression model and advanced to a more sophisticated BERT-based model.

The logistic regression achieved an initial accuracy of 69.56%, demonstrating the model's capability to categorize problems with well-defined mathematical terminologies, particularly evident in categories like Precalculus. However, it struggled with categories featuring broad and fundamental concepts, such as Prealgebra.

Advancing the research, the BERT-based model significantly improved performance, achieving an overall accuracy of 77%. This model demonstrated superior precision, recall, and F1-scores across all categories, validating BERT's efficacy in handling the complex linguistic patterns of mathematical problem statements. Notably, it showed substantial improvement in categories like Number Theory and Intermediate Algebra, indicating its strength in capturing nuanced mathematical expressions.

Despite these achievements, the study identified challenges in uniformly improving performance across all categories, suggesting opportunities for future research. Potential directions include expanding the dataset, refining preprocessing methods, exploring hyperparameter optimization, and employing ensemble techniques to enhance model robustness and accuracy.

In essence, the research confirmed the potential of sophisticated NLP models to significantly enhance automated educational content classification, setting the stage for future advancements in educational technologies.

## 6 References

[1] D. Hendrycks, "Competition Math Dataset," Hugging Face, [Online]. Available: https://huggingface.co/datasets/hendrycks/competition_math. [Accessed: 11, June, 2024].

[2] J. Yao, Z. Zhou, and Q. Wang, "Solving Math Word Problem with Problem Type Classification," in *Proceedings of the Natural Language Processing and Chinese Computing (NLPCC 2023), Lecture Notes in Computer Science*, vol. 14304, pp. 123–134, Springer, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-44699-3_12. [Accessed: 08, October, 2023].