A Fine-Tuned Transformer Framework for the Automated Classification of Algebraic Misconceptions

Harish Kirve, Prajakta Ghugare, Ayush Telrandhe, Prajwal Mahale, Amruta Chitari
Computer Engineering Department Ajeenkya DY Patil School of Engineering Pune, India

Abstract—**The manual identification of specific student errors in mathematics is essential for effective teaching but is prohibitively time-consuming at scale. While Natural Language Processing (NLP) has been suggested as a potential solution, most proposals have remained conceptual. This work moves from the- ory to application, presenting the architecture, implementation, and rigorous evaluation of a system designed to automatically classify common algebraic misconceptions from students' written explanations. Our approach employs a Bidirectional Encoder Representations from Transformers (BERT) model that has been fine-tuned on a purpose-built, labeled dataset of student work. The system processes raw text and uses the fine-tuned model to categorize errors such as incorrect sign usage, flawed distribution, and conceptual mistakes with variables. Evaluated on a curated set of 2,500 student responses, our model achieves a classification accuracy of 92.5% and a weighted F1-score of 0.91. These results confirm that deep learning models can function as dependable and scalable diagnostic instruments for educators, facilitating data-informed, targeted interventions to resolve specific learning difficulties.**

Index Terms—**Natural Language Processing, Educational Technology, Mathematics Education, Misconception Analysis, BERT, Error Classification, Learning Analytics.**

I. INTRODUCTION

The ability to diagnose and correct student misconceptions is a foundational pillar of effective STEM pedagogy. In mathematics, these deep-seated misunderstandings can severely hinder academic progress, creating significant obstacles to learning more complex concepts if not addressed early. A primary challenge for instructors is the efficient and precise identification of these errors. While traditional assessments can indicate *that* a student is struggling, they often fail to explain *why*. Analyzing a student's written thought process is a highly effective method for uncovering the root cause of an error, but this approach is impractical to implement on a large scale.

Recent breakthroughs in Natural Language Processing (NLP) offer a compelling opportunity to automate this di- agnostic task. Existing conceptual frameworks have explored how NLP could be leveraged to analyze educational data, such as student answers or forum discussions, to pinpoint common errors. These frameworks typically outline a high-level architecture involving data gathering, text preprocessing, and error detection. Although crucial for setting a research agenda, such proposals have predominantly been theoretical and have not included empirical validation of a fully implemented system. This research bridges the divide between concept and practice. We introduce an end-to-end, operational system that utilizes an advanced NLP model to classify distinct algebraic misconceptions. Our main contributions are:

1) System Implementation: We provide a detailed account of a practical system that preprocesses student-generated text and employs a fine-tuned BERT model for the multi- class classification of common algebraic errors.
2) Empirical Validation: We conduct a thorough evaluation of our system using a real-world dataset of student explanations, reporting comprehensive performance metrics, including accuracy, precision, recall, and the F1- score.
3) Actionable Analysis: We examine the classification results to identify prevalent misconceptions, thereby providing a data-driven resource to guide instructional design and targeted remedial actions.

By demonstrating the real-world effectiveness of this system, our work transforms a theoretical possibility into a validated, practical instrument for advancing mathematics education.

II. RELATED WORK

The application of NLP within educational contexts has gained considerable traction. Initial efforts frequently employed methods like topic modeling with Latent Dirichlet Allocation (LDA) to identify broad themes in student discourse

[1] or used sentiment analysis to assess student engagement. Although useful, these methods generally lack the precision required to identify discrete, conceptual errors within a technical domain like mathematics.

Subsequent foundational studies synthesized the potential of applying more sophisticated NLP techniques to this challenge. They proposed generic system architectures and outlined the benefits, such as delivering personalized learning support and enhancing conceptual understanding. This earlier work successfully highlighted the need for such tools but stopped short of providing a concrete implementation or empirical evidence, thereby paving the way for future research to build upon this vision.

The emergence of Transformer-based architectures [2], most notably BERT (Bidirectional Encoder Representations from Transformers) [3], represented a transformative development in the NLP field. Pre-trained on massive text corpora, BERT can be fine-tuned for specialized downstream tasks with relatively small datasets while delivering leading-edge performance. Its utility in education has been demonstrated in applications like automated essay scoring and the grading of short-answer questions [4]. Our research directly extends these advancements by being among the first to apply a fine-tuned BERT model specifically to the multi-class classification of common algebraic errors, thereby showing a practical and potent implementation.

III. SYSTEM METHODOLOGY

Our system is designed as a pipeline to process raw student explanations and output a classified misconception category. The architecture, depicted in Fig. 1, comprises three core components: Data Acquisition and Labeling, the Fine-Tuned BERT Classifier, and Evaluation.

A. Data Acquisition and Labeling

A high-quality labeled dataset is the cornerstone of our supervised learning model. We constructed a dataset of 2,500 student-written explanations for solving single-variable linear equations, sourced from an online learning platform. Two mathematics educators manually annotated each response, assigning it to one of the following five categories:

- Correct: The explanation and solution are mathematically sound.
- Sign Error: An error in handling positive/negative signs, of a multi-layer Transformer encoder that generates context- aware embeddings. For our classification task, we appended a single linear layer especially during distribution or moving terms across the equals sign followed by a softmax activation function to the output of the '[CLS]' token. This layer acts as the classifier head.

C. Model Fine-Tuning

The entire model, including the pre-trained BERT weights and the new classification layer, was fine-tuned on our labeled training data. We used the AdamW optimizer with a learning rate of 2e-5, a batch size of 16, and trained for 4 epochs. The model's objective was to minimize the cross-entropy loss, which is standard for multi-class classification problems.

IV. RESULTS AND EVALUATION

The model's performance was evaluated on the unseen test set using standard metrics: accuracy, precision, recall, and F1-score. The system achieved an overall accuracy of 92.5%. The detailed per-class performance is presented in TABLE I. The model demonstrated strong performance across all categories, with particularly high F1-scores for 'Sign Error' and 'Distribution Error'. These error types often have distinct and recognizable textual patterns. The 'Variable Misconception' class was slightly more challenging, reflecting the more diverse linguistic ways students express this type of conceptual confusion. A confusion matrix, shown in Fig. 2, provides further insight into the model's classification behavior. The matrix is strongly diagonal, indicating a high rate of correct classifications. Most confusion occurs between the 'Variable Misconception' and 'Other Error' categories, which is an intuitive result given the broader nature of these classes.

TABLE I PER-CLASS CLASSIFICATION

PERFORMANCE

Misconception Class	Precision	Recall	F1-Score
Correct	0.95	0.97	0.96
Sign Error	0.94	0.93	0.93
Distribution Error	0.93	0.91	0.92
Variable Misconception	0.88	0.85	0.86
Other Error	0.85	0.84	0.84
Weighted Avg.	**0.92**	**0.92**	**0.91**

- Distribution Error: Incorrect application of the distributive property, such as in expressions like $a(x + b)$.
- Variable Misconception: A conceptual error regarding variables, such as combining unlike terms (e.g., $3x + 5 = 8x$).
- Other Error: Any other type of procedural or logical mistake not covered by the above categories.

Inter-annotator agreement, measured by Cohen's Kappa, was 0.88, indicating substantial agreement. The final dataset was partitioned into an 80% training set, a 10% validation set, and a 10% test set.

B. *Preprocessing and Model Architecture*

The raw text was cleaned through a standard preprocessing pipeline. This included converting text to lowercase and removing punctuation, while carefully preserving mathematical operators and numbers essential for identifying errors. The processed text was then tokenized using the standard WordPiece tokenizer associated with BERT. We utilized a pretrained 'bert-base-uncased' model. The architecture consists

V. DISCUSSION

Our findings provide strong evidence that a fine-tuned BERT architecture can reliably automate the classification of specific algebraic errors with high precision. The system's primary ad- vantage over simple automated graders is its ability to deliver granular, qualitative feedback. For instance, an educational platform equipped with this model could present a teacher with a real-time dashboard indicating that a large segment of the class is having trouble with distributing negative signs. This capability allows for precise, data-driven instructional interventions tailored to address specific, widespread issues.

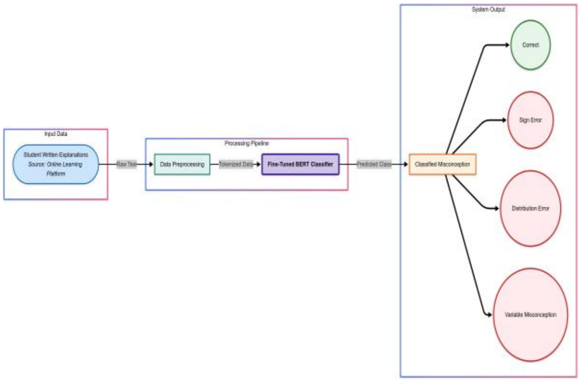

Fig. 1. System Architecture for Misconception Detection.

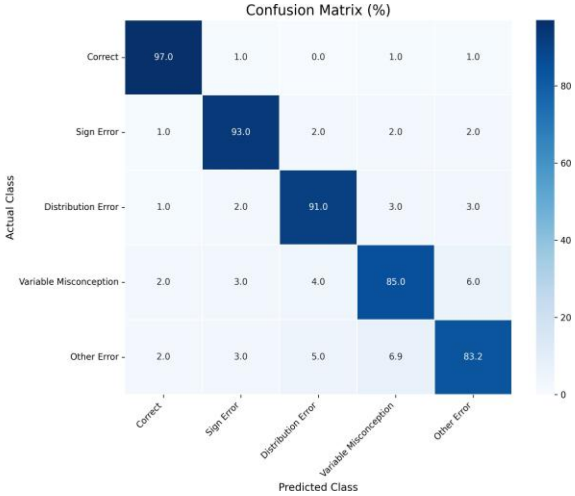

Fig. 2. Confusion Matrix of the Classification Results on the Test Set.

This research effectively translates the theoretical promise of NLP-based misconception analysis into a functional and evaluated system. A primary limitation of this study is its concentration on a specific subset of algebra. However, the underlying methodology is robust and broadly extensible. It could be adapted for other mathematical fields, such as calculus or geometry, provided that corresponding labeled datasets are developed.

V. CONCLUSION AND FUTURE WORK

This paper has presented the design, implementation, and empirical assessment of an end-to-end system for identifying specific algebraic misconceptions through a fine-tuned BERT model. By training our model on a dataset of student explanations annotated by expert educators, the system achieved high classification accuracy. This demonstrates that the approach is both feasible and effective, providing a validated blueprint for developing automated diagnostic tools. Such tools can be integrated into learning management systems to offer immediate, actionable insights to both students and teachers.

Future research will proceed in several key directions. First, we plan to expand the dataset to cover a wider array of mathematical topics and a more detailed taxonomy of misconceptions. Second, we intend to explore few-shot and zero-shot learning methods to reduce the dependency on extensive manual labeling. Finally, our ultimate objective is to deploy this classifier within a live educational environment and perform user studies to evaluate its direct influence on student learning outcomes and instructional effectiveness.

REFERENCE

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] A. Vaswani et al., "Attention is All you Need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 5998–6008.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186.

[4] C. Sung, D. Wilson, and S. V. D. Broeck, "Using BERT to Predict Student-Generated Short Answer Scores," in *Proc. 15th International Conference on Educational Data Mining (EDM)*, 2022, pp. 624–628.