Eedi: Mining Misconceptions in Mathematics

Padmanabh Butala
pb8176@rit.edu

Harshal Chalke
hc4292@rit.edu

***Abstract*—Understanding math misconceptions can be challenging for students, often leading to wrong answers in diagnostic questions. Fixing these misconceptions involves matching wrong answers (distractors) with specific labels, which takes a lot of time and is often inconsistent when done by humans. This project focuses on creating a Natural Language Processing (NLP) model using Machine Learning (ML) to predict the connection between misconceptions and distractors in multiple-choice questions. Using data from Eedi, Vanderbilt University, and The Learning Agency Lab, our goal is to make the labeling process faster, more accurate, and reliable. The model will be tested using the Mean Average Precision @ 25 (MAP@25) score to ensure it works well for both common and new misconceptions. This solution will help teachers tag misconceptions more easily, improving the learning experience for students and making education better for everyone.**

I. Introduction

Misconceptions in mathematics often hinder students from achieving a deeper understanding of key concepts. These misconceptions are frequently reflected in the incorrect answers, or distractors, chosen by students in diagnostic assessments. Tagging these distractors with the corresponding misconceptions can provide valuable insights to educators, enabling them to tailor their teaching strategies. However, manually tagging distractors is not only time-consuming but also prone to inconsistencies. To address this issue, we are motivated to develop automated solutions that can streamline the process while maintaining high accuracy.

In this project, we approach the problem using two distinct methodologies: classical machine learning techniques and large language models (LLMs). Each methodology offers unique strengths and challenges. The classical ML methods focus on interpretable and computationally efficient algorithms, while the LLM approach leverages the power of state-of-the-art models to handle complex semantic relationships.

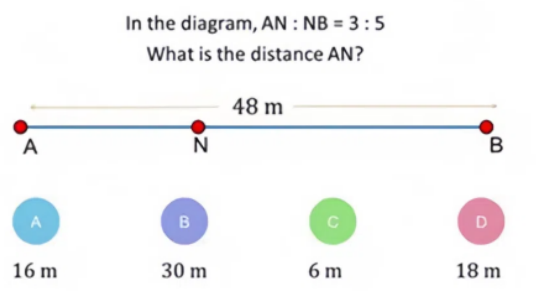
The classical machine learning approach consists of two key methods. The first method involves Latent Semantic Analysis (LSA), which is a statistical technique used to identify patterns in the relationships between terms and concepts in a text corpus. By reducing the dimensionality of the data, LSA enables the model to uncover latent structures that correlate misconceptions with distractors effectively. The second method utilizes word embeddings, where each word or phrase is mapped to a vector in a continuous space. These embeddings capture semantic meaning and relationships, providing a robust foundation for understanding the connections between distractors and misconceptions.

For the LLM approach, we utilize the QWEN 2.5 Instruct model, a cutting-edge language model with 7 billion parameters. This model is designed to excel in instructional tasks, which makes it particularly suited to understand and predict the nuanced relationships between misconceptions and distractors. The model's ability to process both textual and mathematical content allows it to generalize well to a wide range of questions and misconceptions. By fine-tuning this model on our dataset, we aim to achieve high performance in predicting correct misconception labels.

II. Dataset Overview

The dataset used for this project was provided by Eedi, in collaboration with Vanderbilt University and The Learning Agency Lab. It consists of diagnostic multiple-choice questions designed to assess students' understanding of various mathematical constructs. Each question includes one correct answer and three distractors, which are carefully crafted to align with specific misconceptions. This allows the dataset to serve as a valuable resource for studying the relationship between distractors and misconceptions.

Each question in the dataset is associated with a unique identifier, along with metadata such as the subject, construct, and question text. The dataset also includes labels that map each distractor to its corresponding misconception, enabling supervised learning approaches. The text content of the questions and answers was extracted from images using a human-in-the-loop OCR process, ensuring high-quality text data for analysis.



In the diagram, AN : NB = 3 : 5
What is the distance AN?

Fig. 1. Example of a diagnostic question with labeled misconceptions.

An example of a diagnostic question from the dataset is shown in Figure 1. The options are labeled with specific misconceptions as follows:

- **A:** Divides total amount by each side of the ratio instead of dividing by the sum of the parts

- **B:** Mixes up sides of a ratio
- **C:** Finds one part of a ratio but doesn't multiply that by the number of parts needed
- **D:** Correct answer

The training dataset provides both question-answer pairs and their respective misconception labels, while the test dataset requires participants to predict these labels. Additionally, the dataset includes a mapping file that links each misconception ID to its description, providing further context for model development. Evaluation of predictions is performed using the Mean Average Precision @ 25 (MAP@25) metric, which rewards models for ranking relevant misconceptions higher.

This dataset presents unique challenges due to the complexity of mathematical language and the variability in misconceptions. Furthermore, the presence of new or previously unidentified misconceptions in the test data highlights the importance of developing robust models capable of generalizing beyond the training data. These challenges underscore the significance of leveraging both classical ML methods and LLMs as separate approaches to tackle this problem effectively.

III. Methodologies

A. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a statistical technique that identifies relationships between words in a corpus by mapping them into a lower-dimensional semantic space. Below is an explanation of how the LSA-based pipeline works:

- **TF-IDF Vectorization:** The *Term Frequency-Inverse Document Frequency (TF-IDF)* technique is used to represent text data numerically. TF-IDF assigns weights to words based on their importance, reducing the influence of commonly used words while emphasizing rare but significant terms.
- **Truncated Singular Value Decomposition (SVD):** SVD is used to reduce the dimensionality of the TF-IDF matrix, extracting latent features that capture the essential relationships between words and concepts.
- **Cosine Similarity:** After dimensionality reduction, cosine similarity measures the semantic closeness between text representations (e.g., question-answer pairs and misconception descriptions).
- **Evaluation:** The model is evaluated using the Mean Average Precision at 25 (MAP@25), which quantifies how well the top 25 predicted misconceptions align with the ground truth.

This method benefits the problem by providing a structured way to uncover latent relationships in text data, enabling the identification of misconceptions with high interpretability and efficiency. Its reliance on dimensionality reduction also ensures scalability for larger datasets.

B. Word Embedding-Based Approach

Word embeddings represent words as dense vectors in a continuous vector space, capturing semantic and syntactic relationships. This approach uses pre-trained embeddings from GloVe, FastText, and Word2Vec to enhance text representation. Below is an outline of the methodology:

- **Pre-trained Word Embeddings:** Models like GloVe, FastText, and Word2Vec provide pre-trained vectors for a vast vocabulary. These embeddings are used to map words in the dataset into high-dimensional spaces where similar words are closer together.
- **Sentence Embeddings:** Sentence embeddings are generated by averaging the word embeddings of all words in a sentence. This provides a fixed-length representation for variable-length text data, such as question-answer pairs.
- **Dimensionality Reduction:** Using Truncated SVD, the high-dimensional sentence embeddings are reduced to lower dimensions, making computation more efficient and revealing latent patterns in the data.
- **Cosine Similarity:** Similar to LSA, cosine similarity is used to measure the closeness of question-answer pairs and misconceptions in the reduced space.
- **Evaluation:** The MAP@25 metric is used to assess the model's ability to rank relevant misconceptions higher in the predictions.

This method is beneficial as it captures richer semantic and syntactic information from the text, which is particularly useful for nuanced and context-heavy tasks like identifying misconceptions. By leveraging pre-trained embeddings, it can generalize well even on limited datasets, and its integration with dimensionality reduction ensures computational efficiency.

C. Large Language Models (LLMs)

To be continued.

IV. Results

The following table summarizes the MAP@25 scores for different methods:

TABLE I
Evaluation Results for LSA and Word Embedding Methods

Method	**MAP@25 Score**
LSA (50 components)	0.3220
LSA (100 components)	0.3176
LSA (200 components)	0.3163
LSA (300 components)	0.3194
Best LSA (50 components)	**0.3220**
GloVe Embedding	0.3328
FastText Embedding	0.3292
Word2Vec Embedding	0.3246
Combined Embeddings	**0.3358**

A. Steps in the Evaluation

1) **MAP@25:**

$$\text{MAP@25} = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{\min(n,25)} P(k) \times \text{rel}(k)$$

2) **Predicted Misconception IDs:** For each validation sample, the model outputs the top 25 misconception IDs

based on cosine similarity between the reduced embeddings (e.g., LSA components or word embeddings) of the test QA pairs and the misconception mappings.

3) **Actual Misconception IDs:** Each incorrect QA pair in the validation or test set has a ground truth misconception ID provided in the dataset.
4) **Relevance Calculation:** Relevance is calculated as a binary array indicating whether the ground truth misconception ID appears in the top 25 predicted misconception IDs.The relevance is represented as a binary array where: 1 indicates that the true misconception ID is in the predictions. 0 indicates that it is not.

B. Analysis of Results

The results show that LSA achieves its best performance with 50 components (MAP@25 = 0.3220). However, the Word Embedding-Based Approach outperforms LSA, particularly when combining embeddings from multiple models (MAP@25 = 0.3358). This suggests that richer semantic representations from embeddings capture the nuances of the data more effectively.

Despite relatively low MAP@25 scores, these results are still valuable as the task involves predicting up to 25 misconceptions, which is inherently challenging due to the diverse and complex nature of misconceptions. The ability to achieve these scores indicates the methods' capability to identify relevant misconceptions consistently.

V. Discussion

The results obtained from the evaluation indicate significant insights into the performance and applicability of the proposed methods. The LSA method, while simpler and computationally efficient, struggles to capture the more intricate semantic relationships present in the data. Its best performance with 50 components highlights its sensitivity to parameter tuning, and it demonstrates that dimensionality reduction can uncover meaningful latent patterns. However, its limited MAP@25 score reflects its inability to fully leverage the rich context of the question-answer pairs.

The Word Embedding-Based Approach, on the other hand, consistently outperformed LSA, particularly when combining embeddings from GloVe, FastText, and Word2Vec. This improvement underscores the value of pre-trained embeddings in capturing nuanced semantic and syntactic information. Combined embeddings achieve the highest MAP@25 score of 0.3358, demonstrating that integrating multiple embeddings can leverage the strengths of individual models. The robustness and generalizability of this approach make it highly suitable for tasks involving varied and complex textual data.

The relatively modest MAP@25 scores across all methods highlight the inherent difficulty of the problem. Misconceptions in mathematics often involve subtle and context-dependent distinctions, making accurate predictions challenging. Despite this, the achieved scores are valuable as they consistently rank relevant misconceptions higher, aiding human annotators in the labeling process.

The findings suggest that future work should explore hybrid approaches that combine the interpretability of LSA with the semantic richness of embeddings. Additionally, incorporating domain-specific embeddings or fine-tuning existing models on similar datasets may further improve performance. These results pave the way for developing more robust and scalable solutions to address misconceptions in educational settings.

LLM- Harshal