MAP: Charting Student Math Misunderstandings

Project Overview

This project develops a Natural Language Processing (NLP) model to predict students' math misconceptions based on their explanations. The goal is to identify potential math misconceptions that generalize across different problems using the DeBERTa transformer model.

Problem Statement

Students often reveal systematic incorrect ways of thinking (misconceptions) in their mathematical explanations. This project automates the process of identifying these misconceptions to help improve educational feedback and learning outcomes.

Dataset

- **Train data**: 36,696 student responses with 7 features
- **Test data**: 3 samples for submission
- **Key columns**:
  - `QuestionText`: Math problem description
  - `MC_Answer`: Multiple choice answer
  - `StudentExplanation`: Student's reasoning
  - `Category`: Response category (True_Correct, False_Misconception, etc.)
  - `Misconception`: Specific misconception type (26,836 null values handled)

Data Preprocessing

1. **Null Handling**: 26,836 missing values in Misconception column handled by creating `Category:NA` labels
2. **Class Filtering**: Removed rare classes with fewer than 20 samples for robust training

3. **Text Combination**: Combined QuestionText, MC_Answer, and StudentExplanation with [SEP] tokens

4. **Label Encoding**: Converted text labels to numerical format for model training

Model Architecture

- **Base Model**: Microsoft DeBERTa-v3-small

- **Task**: Sequence Classification

- **Number of Classes**: Varies based on filtered classes (typically 30-40)

- **Training**: Hugging Face Trainer with custom training loop

Training Configuration

- **Learning Rate**: 2e-5

- **Batch Size**: 8 per device

- **Epochs**: 3

- **Max Sequence Length**: 256 tokens

- **Validation Split**: 10% of training data

- **Evaluation Metric**: MAP@3 (Mean Average Precision at 3)

Key Features

- **Class Weighting**: Handles imbalanced class distribution

- **Proper Tokenization**: DeBERTa tokenizer with special token handling

- **Early Stopping**: Model selection based on validation performance

- **Top-3 Predictions**: Generates up to 3 misconception predictions per sample

├── MAP_Charting_Student_Math_Misunderstandings.ipynb # Main notebook
├── submission.csv # Prediction output
├── train.csv # Training data

├── test.csv # Test data
└── sample_submission.csv # Submission format

File Structure

Usage

1. Run the complete notebook cell by cell

2. Model automatically downloads competition data via Kaggle API

3. Training process includes validation and model saving

4. Predictions are generated in the correct MAP@3 format

5. Automatic submission to Kaggle competition

Evaluation Metric

MAP@3 (Mean Average Precision at 3):

- Predicts top 3 Category:Misconception combinations

- Score based on rank of correct prediction

- Format: Space-separated predictions per row

Dependencies

```python

transformers

datasets

torch

pandas

numpy

scikit-learn

kaggle
```