Automated Paragraph Classifier

INTRODUCTION

Automated Paragraph Classification is the task of analyzing a paragraph and determining its rhetorical function within a larger essay. This project focuses on building a 7-class classification model capable of categorizing a paragraph into one of the following discourse types:

1. Lead
2. Position
3. Claim
4. Counterclaim
5. Rebuttal
6. Evidence
7. Concluding Statement

This classification system is useful for automated feedback, essay structure analysis, and educational tools. To achieve this, I fine-tuned a custom classification model based on **DeBERTa-v3-Base**, a state-of-the-art transformer model designed for strong performance on natural language understanding tasks.

This report includes data preprocessing, model architecture, training setup, evaluation, results, and limitations.

DATASET

The dataset was taken from Kaggle competition: Feedback Prize – Evaluating Student Learning.

The dataset consisted of student texts classified into one of seven categories mentioned above.

Each row included:

- Input: Text
- Output: A label with the text's rhetorical role

Dataset details:

- Number of texts: 143832
- Number of classes: 7
- Average text length: 44 words

DATA PREPROCESSING

The dataset was already clean enough to tokenize.

Tokenization:

- Tokenizer: DeBERTa-v3-Base
- Max token length: 256
- All texts truncated to 256 tokens

Padding:

- Static padding was intentionally avoided. The dataset was fed into the model using Hugging Face's **DataCollatorWithPadding**, which performs dynamic padding.

Train-Eval Split:

- Standard 80/20 split was applied.
- Evaluation set was used for metric calculation.

MODEL ARCHITECTURE

A custom model was build using pytorch library.

- Base model: DeBERTa-v3-base
- Output dimension: 7 (classes)
- Pooling: A custom pooler was built using pytorch for attention pooling.
- Loss function: The model uses weighted Cross-EntropyLoss. Class weights were computed with scikit-learn's **compute_class_weight("balanced")** and passed into the loss function to counteract class imbalance.
- Evaluation metrics: Accuracy, F1 (Macro)

TRAINING

Model was trained using Hugging Face Trainer.

Environment:

- Trained on Kaggle GPU T4 x 2
- Total training time: 5 hours

HYPERPARAMETERS

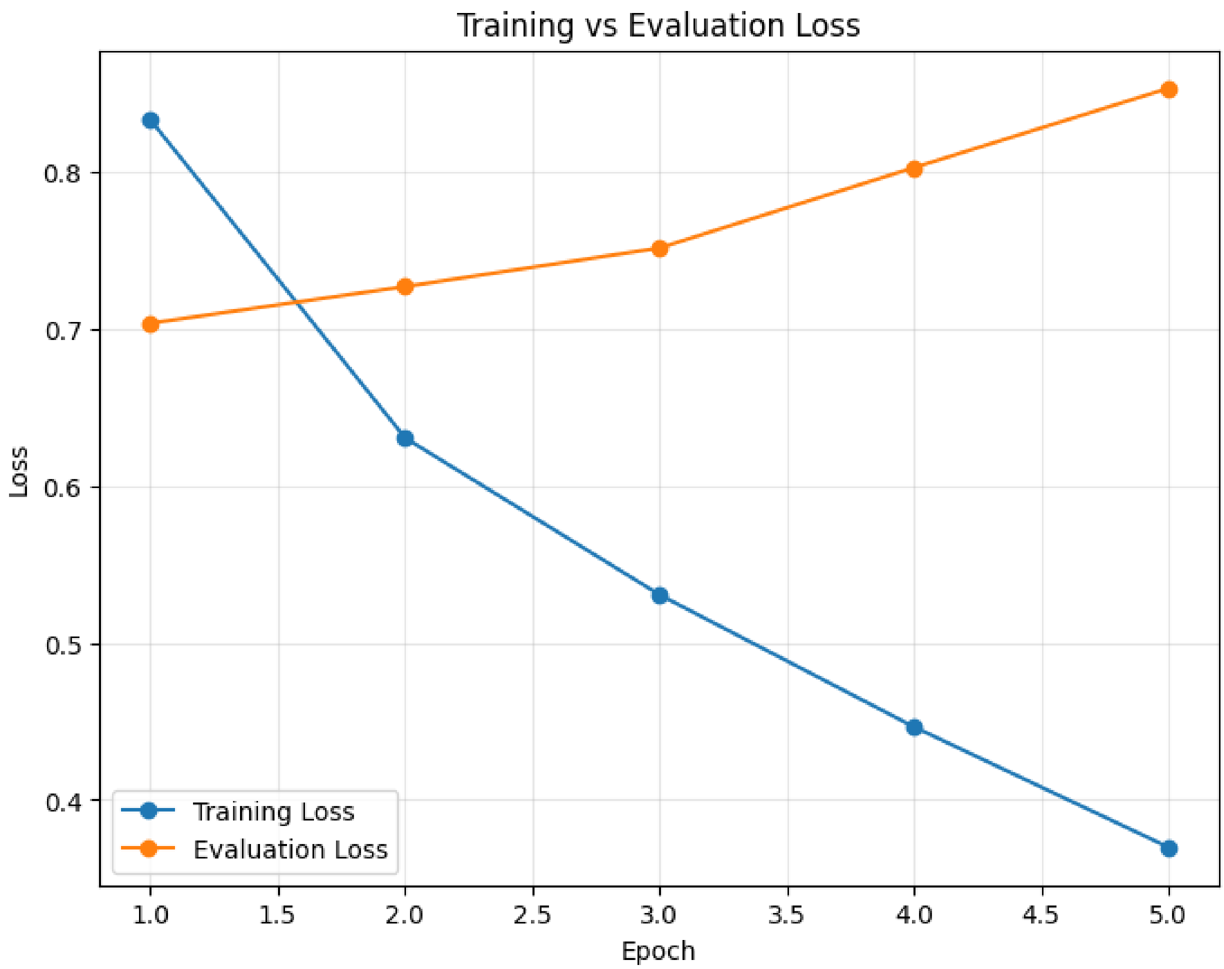
Learning Rate	1e-5
Optimizer	AdamW
Training Batch Size	16
Evaluation Batch Size	16
Weight Decay	0.02
Best Model Evaluation Metric	F1 (Macro)
Epochs	10
Early Stopping	Enabled (patience = 2)

Training stopped at 5^{th} epoch.

EVALUATION

Evaluation table is available is “results” folder with filename “results.csv”.

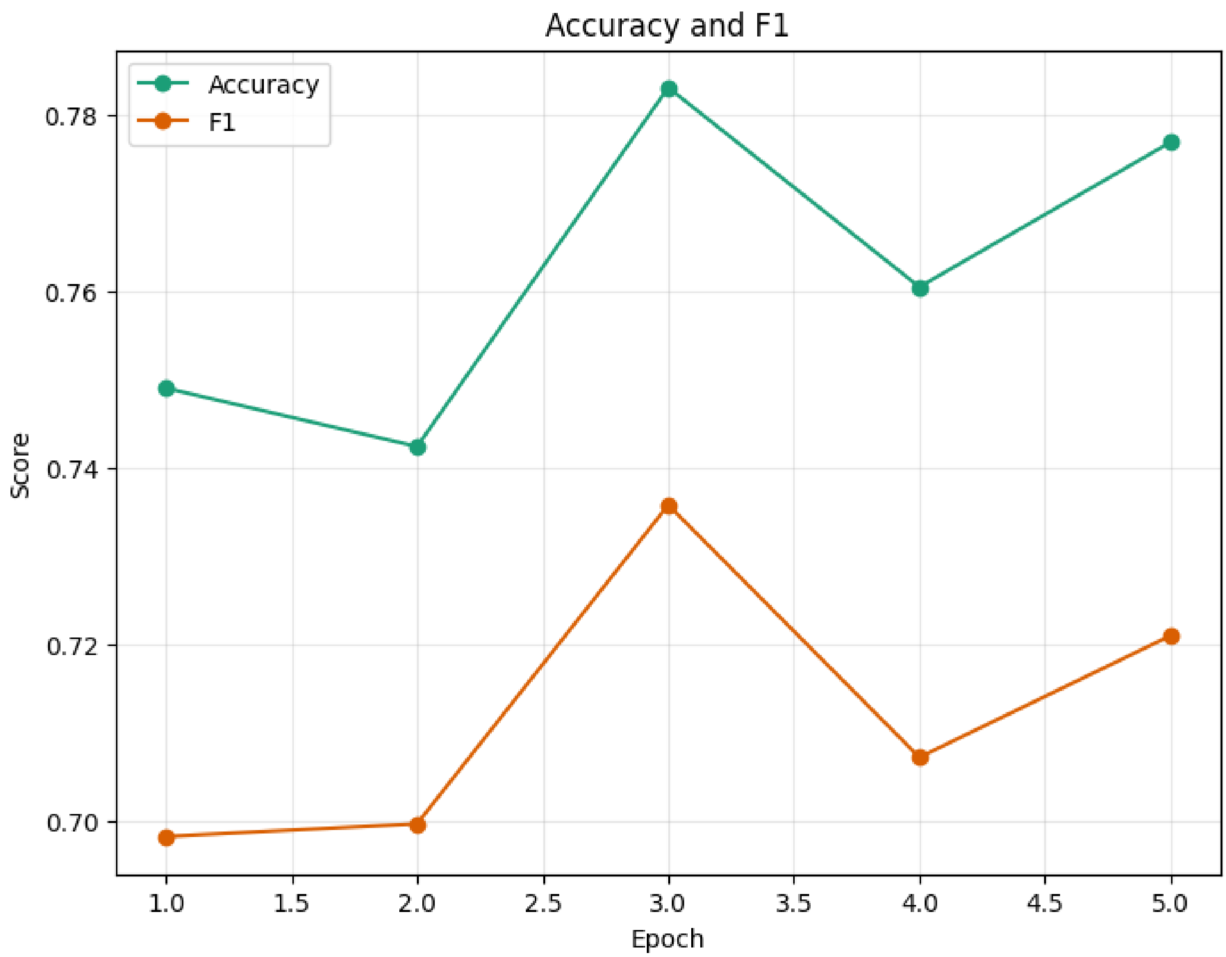
Training vs Evaluation loss:



Observations:

- Training loss decreases across all 5 epochs.
- Slight increase in evaluation loss across epochs.
- Model tends to overfit.

Accuracy and F1 Scores:



Observations:

- Both metrics improved from epoch 1-3.
- Best performance at epoch 3.
- F1-Macro decreasing at later epochs, indicating overfitting.

Final Performance:

- Accuracy: 0.78
- F1 (Macro): 0.73

LIMITATIONS

- Class imbalance: Some categories are underrepresented.
- Domain bias: Performance may drop for texts beyond training domains.
- No explanation: Model cannot provide explanation for its classification.
- Multi-label: Some texts may serve multiple functions (e.g.: Lead + Claim). This model can only classify into one.
- Length: Works best when input is less than 100 tokens.

CONCLUSION

A DeBERTa-v3-based classifier was successfully trained to identify paragraph roles across 7 discourse categories.

With an F1-macro of 0.73, the model demonstrates strong performance for essay structure analysis and educational AI application.