Automated Essay Scoring 2.0: Technical Report

Introduction

This study details the development of an automated essay scoring system for the Learning Agency Lab's Automated Essay Scoring 2.0 competition. The solution employs a fine-tuned DeBERTa-v3-small transformer model with a regression approach, achieving a quadratic weighted kappa (QWK) score of 0.8244 through 5-fold cross-validation. The model successfully demonstrates the potential for reliable, open-source automated essay grading systems that can supplement educator efforts and provide timely feedback to students, particularly in underserved communities where grading resources may be limited.

Background and Problem Statement

Essay writing remains a crucial method for evaluating student learning and performance; however, the time-intensive nature of manual grading creates significant challenges for educators. While Automated Writing Evaluation (AWE) systems offer a potential solution, their high costs have limited widespread adoption, particularly in underserved communities.

The Learning Agency Lab's Automated Essay Scoring 2.0 competition aims to address these limitations by developing improved, open-source essay scoring algorithms that can:

1. Reduce the time and expense associated with manual grading
2. Enable more frequent use of essays in educational assessment
3. Provide students with regular and timely feedback on their writing
4. Make these tools accessible to all educational communities

This competition builds upon the original Automated Student Assessment Prize (ASAP) from 2012 by utilizing a more expansive dataset that includes high-quality, realistic classroom writing samples across diverse economic and geographic populations.

Methodology

Dataset Characteristics

The competition provides what is described as "the largest open-access writing dataset aligned to current standards for student-appropriate assessments." While specific details about the dataset size aren't provided in the code, we can see that the training process involves

approximately 17,307 essays (split across 5 folds with approximately 13,845 training and 3,462 validation samples per fold).

Model Selection

The solution employs Microsoft's DeBERTa-v3-small model, a transformer-based architecture known for its strong performance on natural language understanding tasks. DeBERTa (Decoding-enhanced BERT with disentangled attention) improves upon BERT through:

- Enhanced disentangled attention mechanism
- Improved position encoding
- Better handling of long-range dependencies in text

Scoring Approach

The solution adopts a regression-based approach rather than classification:

- **Regression**: Directly predicts a continuous score value which is then rounded to the nearest integer
- This approach allows the model to better understand the ordinal nature of essay scores and the relative differences between score points

Evaluation Metric

The competition uses Quadratic Weighted Kappa (QWK) as the primary evaluation metric, which:

- Measures agreement between human and machine scores
- Penalizes disagreements based on their squared distance (larger scoring discrepancies are penalized more heavily)
- Ranges from 0 (random agreement) to 1 (complete agreement), though it can go below 0 in cases of worse-than-random agreement

Implementation Details

Model Configuration

- **Base Model**: DeBERTa-v3-small
- **Approach**: Regression (predicting continuous scores rather than discrete classes)
- **Score Range**: Original scores (1-6) converted to labels (0-5) for training, then converted back for final predictions
- **Cross-validation**: 5-fold stratified cross-validation
- **Maximum Sequence Length**: 1024 tokens

- **Special Token Handling**: Added custom tokens for newlines and double spaces, which are important formatting elements in essay assessment

Training Parameters

- **Learning Rate**: 1e-5
- **Batch Size**: 4 (training), 8 (evaluation)
- **Training Epochs**: 4
- **Weight Decay**: 0.01
- **Optimizer**: AdamW with linear learning rate schedule
- **Mixed Precision Training**: Enabled (FP16)
- **Early Stopping**: Based on QWK metric

Text Preprocessing

The implementation includes careful handling of text features important for essay evaluation:

- Addition of special tokens for newlines and double spaces
- Preservation of essay structure through appropriate tokenization
- Truncation to 1024 tokens for overly long essays

Results and Evaluation

Cross-Validation Performance

The model achieved consistent performance across all five folds:

Fold	Final QWK Score
1	0.827
2	0.832
3	0.813
4	0.821
5	0.817

Overall Cross-Validation QWK Score: 0.8244

This score represents strong agreement between the model's predictions and human scores, indicating that the system can reliably grade student essays in a manner consistent with human evaluators.

Training Progression

Each fold showed consistent improvement in the QWK score throughout the training process:

1. **Fold 1**: QWK improved from 0.780 (Epoch 1) to 0.827 (Epoch 4)
2. **Fold 2**: QWK improved from 0.759 (Epoch 1) to 0.832 (Epoch 4)
3. **Fold 3**: QWK improved from 0.786 (Epoch 1) to 0.813 (Epoch 4)
4. **Fold 4**: QWK improved from 0.781 (Epoch 1) to 0.821 (Epoch 4)
5. **Fold 5**: QWK improved from 0.701 (Epoch 1) to 0.817 (Epoch 4)

Training loss consistently decreased across all folds, indicating proper model convergence without apparent overfitting.

Score Distribution

While detailed score distribution analysis isn't provided in the code output, the confusion matrices generated for each fold (though not visible in the text output) would show the distribution of predictions across the six possible score points, highlighting any systematic biases in the model's predictions.

Conclusion

Key Achievements

1. Successfully developed an automated essay scoring system with strong performance (QWK = 0.8244)
2. Implemented a robust cross-validation approach to ensure generalizability
3. Utilized state-of-the-art transformer architecture (DeBERTa-v3-small) for essay understanding
4. Created an ensemble prediction approach by averaging predictions across five model folds

Implications

This work demonstrates the viability of open-source automated essay scoring systems that can:

- Provide reliable, consistent scoring comparable to human evaluators
- Reduce grading burden on educators
- Enable more frequent essay assignments and feedback
- Support educational assessment in underserved communities

Acknowledgement

This notebook is inspired by the work of Hashido Yuto and Chris Deotte.