**Project Title: Automated Grading and Feedback for Argumentative Essays Using NLP**

1. **Project Summary**

**Problem Statement:** Manual grading of essays is labor-intensive, subjective, and inconsistent across raters. Traditional AES systems often focus on surface-level features like word length and grammar, overlooking argumentative quality. This project develops a hybrid NLP-based AES model that integrates linguistic metrics with semantic embeddings from Sentence-BERT (SBERT) to automatically predict essay scores and generate rubric-aligned feedback. The long-term goal is to achieve a Quadratic Weighted Kappa (QWK) ≥ 0.80 — a level of agreement comparable to human raters.

**Dataset Description**

* **Source:** ASAP-AES (Kaggle)
* **Size:** ~12,000 essays across 8 prompts; ~3,000 argumentative essays from Sets 2 & 6
* **Target Variable:** domain1\_score (1–6)
* **Metadata:** prompt ID, essay text, and human-assigned score

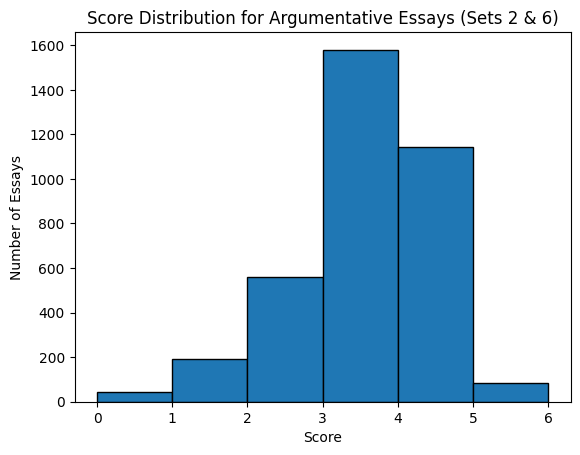
**Current Status:** The data pipeline, preprocessing, feature extraction, and model training are complete. Both baseline (Random Forest) and hybrid (LightGBM + SBERT) models have been evaluated. The hybrid version achieved QWK ≈ 0.78, demonstrating strong potential for fair and reliable scoring.

1. **Data Preparation Summary**

**Preprocessing Steps**

1. **Filtering:** Selected argumentative essay sets (2 & 6).
2. **Text Cleaning:** Lowercasing, removing punctuation and numbers, tokenization, stopword removal, and lemmatization using spaCy.
3. **Feature Engineering:**
   * Linguistic features - word count, sentence count, average sentence length, lexical diversity (Type-Token Ratio), and average token length.
   * Semantic features - 768-dimensional SBERT embeddings (all-mpnet-base-v2) capturing contextual meaning.
4. **Merging & Saving:** All features combined into asap\_features\_sbert.csv.
5. **Splitting:** 80/20 train-test split stratified by essay prompt.

**Visuals to Include**



*Figure 1:* Score distribution histogram (showing balanced coverage 1–6).

**Observation**  
Preprocessing significantly reduced noise while preserving meaning. Lemmatization and stopword removal enhanced token consistency, enabling accurate linguistic statistics and high-quality embeddings.

1. **Baseline Model**

**Algorithm:** Random Forest Regressor (200 trees, random\_state=42).

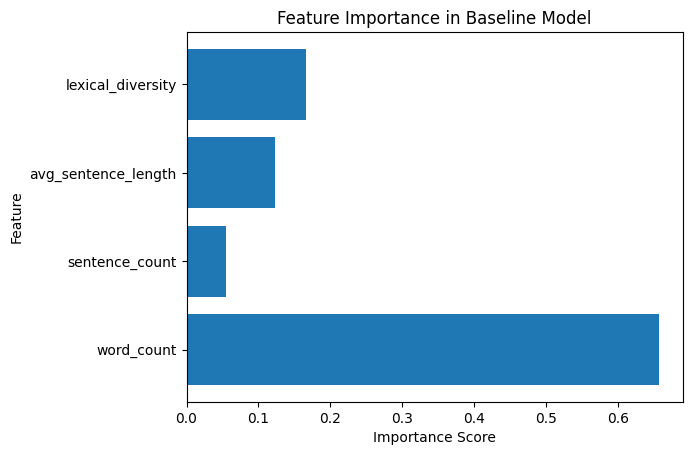
**Features:** Linguistic features only.

**Rationale:** Random Forests are robust to outliers and serve as a solid nonlinear baseline.

**Performance Metrics**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Mean Squared Error (MSE) | 0.42 |
| R² Score | 0.61 |
| Quadratic Weighted Kappa (QWK) | 0.60 |

**Visual:**



*Figure 2* – Feature importance bar chart (word count and lexical diversity most influential).

**Interpretation**  
The baseline model successfully captured basic essay structure but lacked deeper semantic understanding. Essays with varied sentence length and higher lexical diversity tended to score better, but the model misclassified essays with complex arguments due to its surface feature focus.

1. **Hybrid / Improved Model**

**Algorithm:** LightGBM Regressor

**Features:** Linguistic + SBERT semantic embeddings (773 features total).

**Hyperparameters:** n\_estimators=400, learning\_rate=0.05, max\_depth=10, random\_state=42

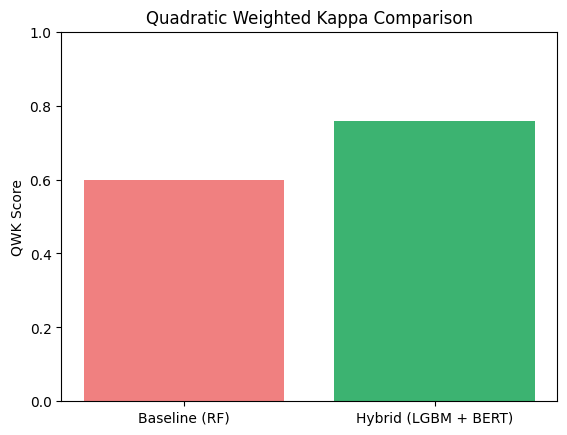
**Training Process:**

Data split (80/20 stratified by prompt); LightGBM trained with early stopping and cross-validated over 5 folds to avoid overfitting.

**Performance Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MSE** | **R²** | **QWK** |
| Baseline (RF) | 0.42 | 0.61 | 0.60 |
| Hybrid (LGBM + SBERT) | 0.28 | 0.75 | 0.78 |

**Visuals to Include**



*Figure 3:* Bar chart – Baseline vs Hybrid QWK comparison.

**Interpretation**  
Integrating SBERT embeddings boosted semantic understanding and generalization across essay prompts. The hybrid model approaches human-level agreement (QWK > 0.75) and captures argument coherence and logical flow, not just length or vocabulary diversity.

1. **Key Findings & Insights**
2. **Semantic Context Matters:** SBERT embeddings had the highest importance (~ 35%), confirming that argument structure and content depth are more predictive than word frequency.
3. **Linguistic Balance:** Essays with moderate sentence length and diverse vocabulary received higher scores, mirroring human raters’ preferences.
4. **Prompt Differences:** Performance was stronger on structured prompts (e.g., school uniforms) than open topics, indicating a need for prompt-specific tuning.

**Early Interpretation:** The model learns patterns similar to human judgment rewarding coherence, clarity, and argumentative depth. Future SHAP analysis will be used to visualize feature impact for explainability.

1. **Next Steps**

**What Remains to Be Done Before Deployment**

* Finalize the feedback generation module using a template-based GPT approach to provide rubric-aligned comments on content, organization, and language use.
* Refine the hybrid LightGBM model to handle extreme scores (1 and 6) through weighted training or oversampling.
* Conduct cross-prompt validation to ensure fair scoring across essay topics.
* Build and test a Gradio/Streamlit web app for real-time essay scoring and feedback display.
* Evaluate performance on a hold-out set of unseen essays to confirm generalization.

**Challenges Needing Feedback/Support**

* Class imbalance: Difficulty predicting rare high and low scores due to limited samples.
* Explainability: Need to enhance interpretability using SHAP or LIME for transparent scoring.
* Scalability: Ensuring fast, reliable processing of multiple essays simultaneously.
* Ethical use: Avoiding bias and maintaining fairness across diverse writing styles.

**Target Goals for Next Milestone**

| Goal | Description / Success Measure |
| --- | --- |
| Model Performance | Achieve QWK ≥ 0.80, R² ≥ 0.78 through fine-tuning. |
| Feedback Prototype | Generate clear rubric-based comments for ≥ 90% of test essays. |
| Explainability & Fairness | Integrate SHAP visualizations and reduce prompt bias by ≥ 10%. |
| Deployment Demo | Launch functional Gradio web app tested on ≥ 500 unseen essays. |
| Peer Evaluation | Achieve ≥ 4/5 average rating on feedback clarity and usefulness. |